Naval Submarine Medical Research Laboratory

NSMRL/50709/TR--2011-0031 **03 October, 2011**

STUDIES ON A SPATIALIZED AUDIO INTERFACE FOR SONAR

by

Thomas P. Santoro
NSMRL

Gregory H. Wakefield
University of Michigan

Agnieszka Roginska
New York University

Approved and Released by:
P. Kelleher, CAPT, MC, USN
Commanding Officer
NAVSUBMEDRSCHLAB

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
03-10-2011	Technical Report	December 2007 - September 2010

4. TITLE AND SUBTITLE
STUDIES ON A SPATIALIZED AUDIO INTERFACE FOR SONAR

5a. CONTRACT NUMBER
N0001409WR20037

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER
0603236N

6. AUTHOR(S)
Thomas P. Santoro, NSMRL

Gregory H. Wakefield, University of Michigan

Agnieszka Roginska, New York University

5d. PROJECT NUMBER
02915

5e. TASK NUMBER
09PRO1005-00

5f. WORK UNIT NUMBER
50709

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Naval Submarine Medical Research Laboratory
Box 900
Groton, CT 06349-5900

8. PERFORMING ORGANIZATION REPORT NUMBER
NSMRL/50709/TR--2011-

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research N0001409WR20037

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

This report summarizes an investigation of the application of a spatialized audio interface over headphones for use in submarine sonar operations and training that was performed in the period from December 2007 through September 2010 as part of a larger ONR study led by the Johns Hopkins University Applied Research Laboratory with engineering support from the Naval Undersea Warfare Center, Newport. NSMRL, with collaborators at University of Michigan and New York University, provided applied binaural research on human auditory processing of spatialized (or 3D) audio over headphones. A comprehensive summary is presented of a series of experiments in six focus areas quantifying human listener performance as a function of variations in critical parameters of this new audio interface. Recomendations are also made for further study in four of the focus areas where significant performance improvement may be gained.

15. SUBJECT TERMS
spatialized audio interface, 3D audio, audio interface

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			NAVSUBMEDRSCHLAB Commanding Officer
U	U	U	SAR	83	19b. TELEPHONE NUMBER *(Include area code)* 860-694-3263

Reset

Standard Form 298 (Rev. 8/98)
Prescribed by ANSI Std. Z39.18

[THIS PAGE INTENTIONALLY LEFT BLANK]

STUDIES ON A SPATIALIZED AUDIO INTERFACE FOR SONAR

Author(s)

Thomas P. Santoro
NSMRL

Gregory H. Wakefield
University of Michigan

Agnieszka Roginska
New York University

Naval Submarine Medical Research Laboratory

Approved and Released by:

CAPT P. Kelleher, MC, USN
Commanding Officer
Naval Submarine Medical Research Laboratory
Submarine Base New London Box 900
Groton, CT 06349-5900

ADMINISTRATIVE INFORMATION

This work was supported by Work Unit 50709.

[THIS PAGE INTENTIONALLY LEFT BLANK]

ABSTRACT

This report summarizes an investigation of the application of a spatialized audio interface over headphones for use in submarine sonar operations and training that was performed in the period from December 2007 through September 2010 as part of a larger ONR study led by the Johns Hopkins University Applied Research Laboratory with engineering support from the Naval Undersea Warfare Center, Newport. NSMRL, with collaborators at University of Michigan and New York University, provided applied binaural research on human auditory processing of spatialized (or 3D) audio over headphones. A comprehensive summary is presented of a series of experiments in six focus areas quantifying human listener performance as a function of variations in critical parameters of this new audio interface. Recomendations are also made for further study in four of the focus areas where significant performance improvement may be gained.

[THIS PAGE INTENTIONALLY LEFT BLANK]

CONTENTS

ACKNOWLEDGEMENT

The authors wish to acknowledge Ms. Margaret Beecher, JHU/APL, for her enduring leadership, guidance, and inspiration throughout this program.

INTRODUCTION

The addition of spatialized audio to visual displays for sonar is much akin to the development of talking movies in the early days of cinema and can be expected to have as profound an effect on the activity. The presention of a spatialized auditory surround with corresponding visual displays is intended to promote observer "suspension of disbelief" in the combined synthetic environment and fully engage perception and cognition in the tactical situation at hand. In this study, NSMRL collaborated with Dr. Gregory Wakefield of the University of Michigan and Agnieszka Roginska of New York University to address the following issues:

- The restoration of human audition in the sonar surveilance task to its natural sensory role in coordination with vision in normal human perception.
- The implementation of spatialized sonar audio to simulate the natural 3D listening environment in which the human auditory system is accustomed to function and achieve its optimal sensory performance.
- The tailoring of new sonar signal processing functions to render an optimized auditory presentation for human listening.

The scientific investigation was based on the following hypothesis:

- A spatialized audio compliment to the visual sonar interface will provide operators with the capability to simultaneously monitor acoustic events at all locations in the undersea surround resulting in increased opportunities for detection, recognition and tracking, quicker bearing estimation, and expanded situation awareness.
- Naturalistic coupling of the visual/aural senses will have a synergistic effect allowing each sense to magnify the power of the other without additional cognitive load or operator fatigue.
- Advanced signal processing techniques for audio will be successful in rendering and "de-cluttering" noisey, multi-sound displays just as complex visual imagery can be enhanced by techniques that facilitate optimal information transfer to the observer.

In the three years of the study, five major experiments on basic auditory behaviors were completed. In addition, extensive exploration of the auditory "split-channel" or "center-surround" combined monaural - binaural auditory display was completed. An examination of alternative auditory search techniques for navigation in a virtual acoustic surround was performed. An advanced audio signal processing technique was adapted to optimize spatial unmasking for sonar sounds. Head Related Transfer Function (HRTF) rendering techniques capable of handling a very large number of sources were developed. A collection of over 30 HRTF data sets for sonar operators from SUBASE NLON were measured in the NSMRL anechoic chamber facility.

The measured HRTFs were used by JHU/APL collaborators to study operator performance in test scenarios with spatialized audio on a sonar simulator provided by NUWC. A large sound proof room in the NSMRL auditory suite was secured by SUBASE NLON Public Works for the APL operational tests with classified sonar data. It was partitioned, sound proofed, and air conditioned to house the NUWC sonar simulator and a 3D audio rendering engine and to suppress their operating sounds from the main human subject test area. A description of the technical support by the NSMRL team that provided for the experiments in the new facility to familiarize users with the fundamentals of audio perception in three spaces and to evaluate users' inherent listening capabilities prior to the operational testing follows.

SUPPORT FOR OPERATIONAL SONAR HUMAN PERFORMANCE TESTING

In addition to soliciting qualified sonar operators from the SUBASE NLON community to perform as subjects in the operational sonar testing performed by the program's lead lab, JHU/APL, and providing a secure, sound-proof experiment space for the experiments, the NSMRL team also developed and supported various basic audio-visual interfaces used to introduce 3D audio to subjects and evaluate their initial impressions and capabilities with this new interface. A major accomplishment in this regard was the development of a streaming audio capability that allowed the integration of real-time audio support into Matlab. Well-known for its promotion of rapid prototyping in signal processing, Matlab has been an ideal tool for studying all of the various aspects of spatial audio and binaural hearing except for the fact that it had been impossible to play audio continuously and allow users to adjust the parameters of the system generating the audio. By developing a streaming audio capability, we were able to expand the capabilities of our tests in ways that would have required much more time to program in other software environments.

After going through several audio development projects, we were able to isolate the primary functions that a spatial-audio development system should support in Matlab. By isolating these into separate, callable GUI objects, we created what we call the *Audition Test Bed*. As diagrammed in Figure 1 below, this interface allows the experimenter to program the main application to support their own psychophysical procedure or behavioral testing paradigm. A prototype application includes all the necessary function calls to initialize and communicate with objects that support streaming audio, audio file i/o, spatial rendering, and data logging of user response. Using this prototype, the experiments in granularization, auditory search, and complexity reduction were built.

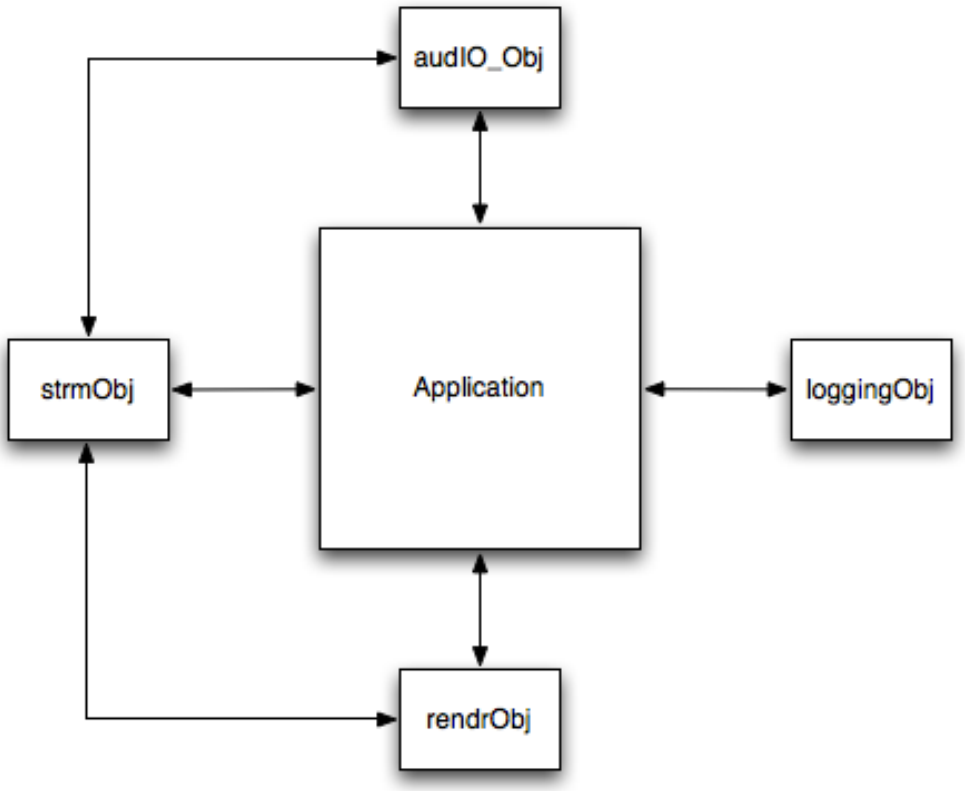

Figure 1. The Audition Test Bed is built in Matlab around a core application and stand-alone supporting objects. The latter provide audio streaming, audio file i/o, spatial rendering, and user-response logging. The suite of objects has permitted the rapid development and refinement of a number of psychophysical experiments, including the four basic spatial audio orientation modules of the Human Performance Measurement team.

This packaging of capabilities in Matlab also created opportunities to help the Human Performance Testing program while staying within budget. As the team began piloting their various studies, it became clear that sonar operators would need some form of training in spatial audio before they could take full advantage of the displays. Working with the team, four orientation modules were piloted and refined. In all cases, specialized versions of the application were refined to meet the needs of the team and capitalize on a number of their training insights.
Highlights of these modules include:

Spatialization

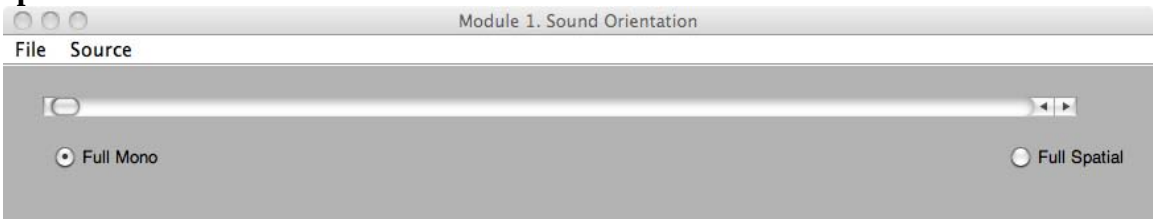

Figure 2. Spatialization GUI. Supports the comparison of mono and spatial rendering of experimenter-selected sources either in binary mode (using the radio buttons) or in a mixed mode (using the slider to control the transition from mono to spatial).

Auditory Perception

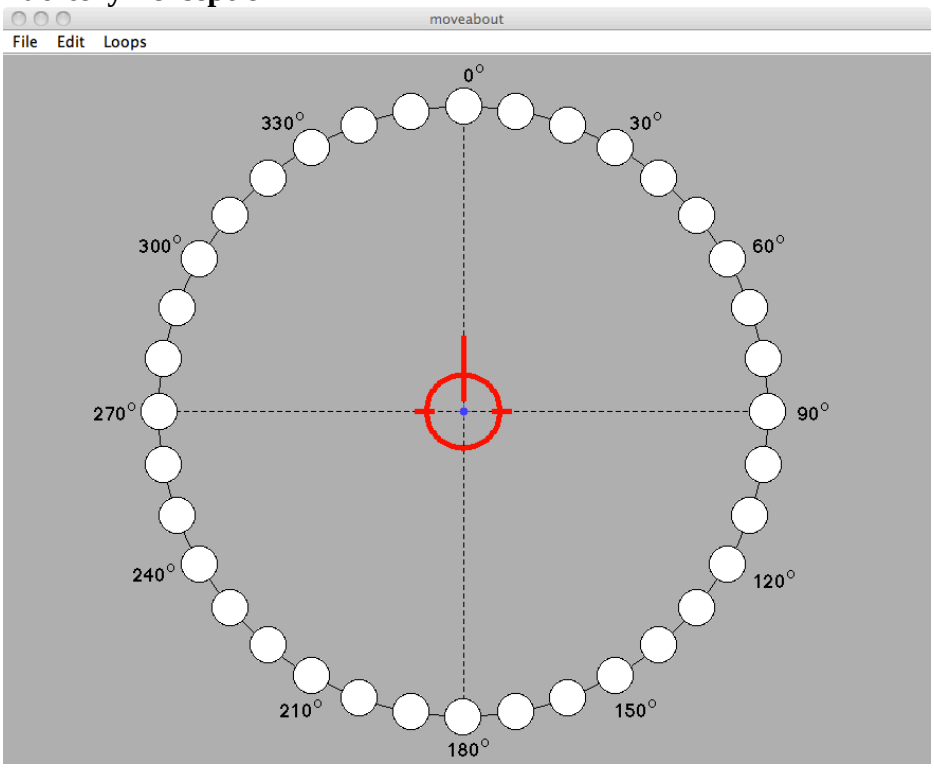

Figure 3. Auditory Perception GUI. Introduces the operator to a top-down perspective of their spatial audio environment using a visual icon and a "circle-of-source-locations". Supports source motion through the selection of various "loops". These loops present sources moving through various hemi-circle and full circle paths around the operator.

Direction Finding

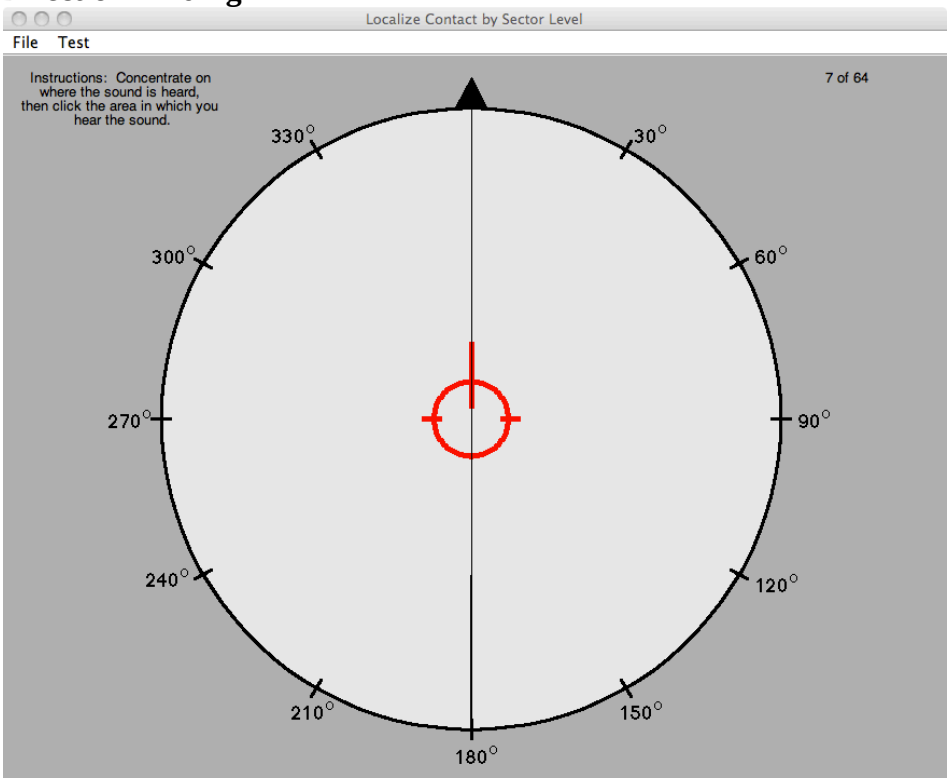

Figure 4. Direction Finding GUI. Supports the orienting of audio location to visual location through the selection of sectors within which a single source is presented. The screen can be divided into hemicircles of various orientations (one example shown above), quadrants, or octants. Testing allows the operator and experimenter to determine how well the operator places sources in general locations of their auditory space.

Multicontact Detection

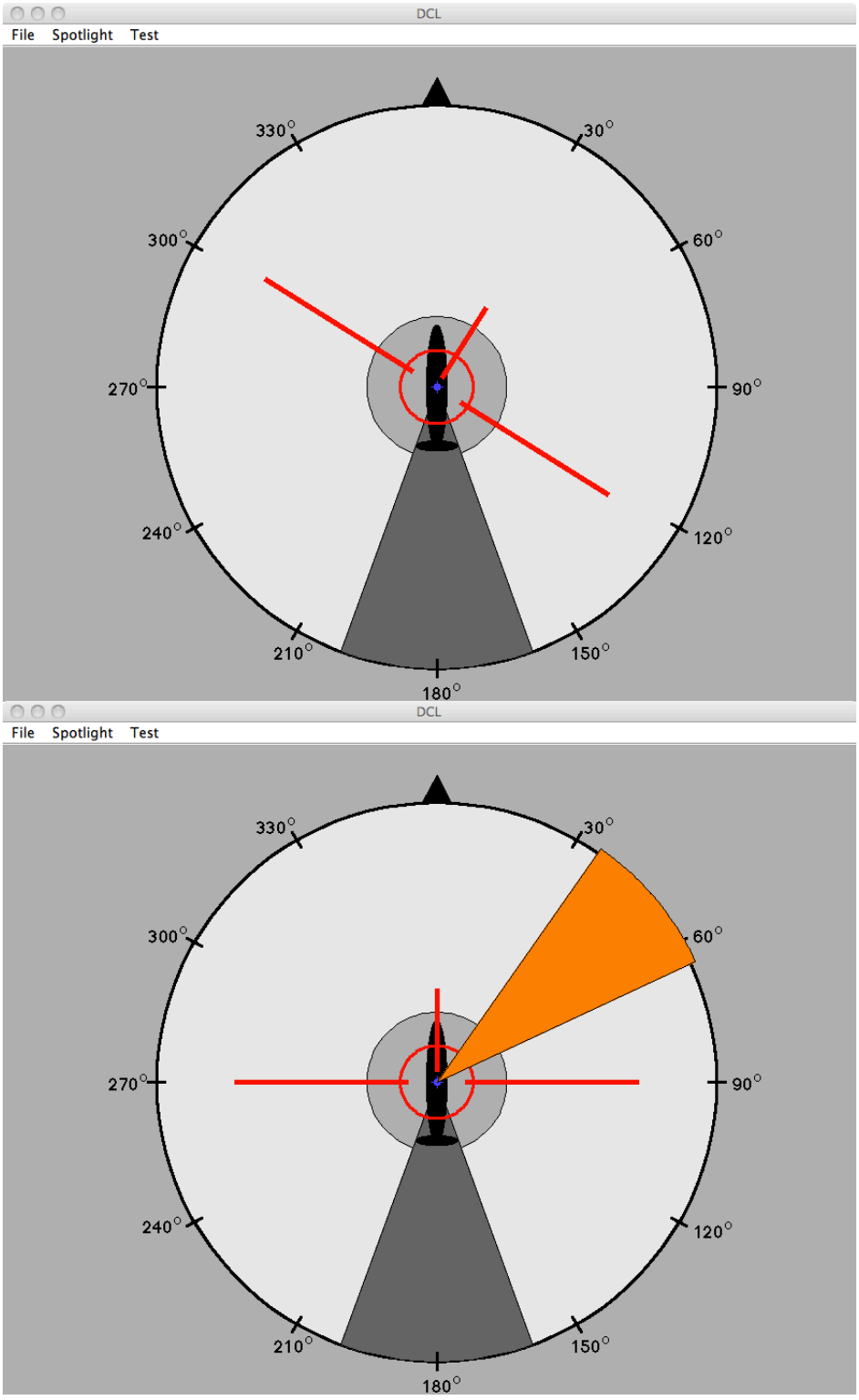

Figure 5. Multiple Contact Detection GUIs (top and bottom). Supports the detection and classification of up to 8 sources presented simultaneously. Orients the operator to the use of head rotation to resolve ambiguities in the spatial display. Introduces the operator to spotlighting – a mixing of 360-degree awareness with single-beam focus.

TOPICS FOR FURTHER INVESTIGATION

In the three years of this study, the NSMRL activity focused on six target areas: (1) listening under conditions of signal, masker, and spatial uncertainty; (2) navigation in a virtual auditory environment; (3) HRTF measurements and selection; (4) improved perceptual signal-to-noise ratio in spatial audio; (5) reduced computational complexity of large multi-input spatial audio systems; and (6) dual task decision-making with single sensory modality information sources. The methods, results, and discussion for each of these areas are presented in the remainder of this report. In addition, recomendations are made on four areas where we feel significant performance improvement remains to be reached through further investigation. These areas are: (1) optimizing the spatial audio immersive experience through the selection of appropriate HRTFs, (2) supporting dynamic binaural cues through hand-operated user interfaces, (3) improving the perceptual signal-to-noise ratio in spatial audio for underwater applications, and (4) reducing the computational complexity of large multi-input spatial audio systems.

The authors strongly recommend that new work on any of the four areas be undertaken within the framework of spatial understanding, cognition, and reasoning. Investigators within the traditional binaural research community have attended primarily to sensory-based questions of detection and localization. Their results point to the limitations of any spatial-audio system, but they do not address how such systems support more cognitively-based tasks. Any future work in any of the four areas above should be clearly directed towards integrating the vast amount of research that already exists in the visual, haptic (tactile), computer and geographical sciences and adopt experimental paradigms that truly address the concerns of the human-system interface for underwater acoustic environments.

LISTENING UNDER CONDITIONS OF SIGNAL, MASKER, AND SPATIAL UNCERTAINTY

Methods

Stimulus design

Rather than drawing upon recordings of surface ships or other examples of signals dominated by rotational machinery noise, we chose a model to generate exemplars of such signals. We did this for three reasons. Firstly, the number of such recordings in the unclassified domain is relatively small. Working within this set was likely to result in over-learned stimuli, which would defeat the purpose of testing for otherwise unknown signals. Secondly, any given recording contains a number of spectral and temporal features that give rise to its particular “signature”. Much is known about binaural masking of such known features. Our desire was to avoid using pre-determined features so that performance would reflect solely the more primitive “periodicity detection” mechanisms in auditory perception which, we hypothesize, would be the first to indicate the emergence of an otherwise unknown source in the environment. Thirdly, such

recordings also contain sea-state noise and other contaminants, which could be interpreted by listeners as alternative cues.

Infrapitch noise signals were originally constructed by Warren (1981) to study primitive "periodicity detection" mechanisms in auditory perception. They consist of a sample of noise that is repeated over and over again. Warren observed that such noises capture the intrinsic nature of machinery noise for repetition rates on the order of 1-10 Hz by producing "whirr", "rumble", "whooshing", and other characteristic attributes. Warren also demonstrated that this perception takes a period of time to build up, but once heard for a particular sample of noise, is quite robust.

Our experiments utilized 1-second infrapitch noises to model rotational machinery-noise sources and pink noise to model sea-state noise. The infrapitch noises were drawn from the same pink noise as the sea-state noise, to eliminate any long-term spectral differences that listeners could use to detect the infrapitch noises. The period of the infrapitch noise was 200 ms, so that the listener heard five repetitions of the pink noise sample during each observation. This provided sufficient exposure for the infrapitch noise to be clearly identified when presented in pink noise alone without any other sources.

Psychophysical procedure

We employed a standard two-interval forced choice adaptive procedure for measuring signal detection threshold (Levitt, 1971). On any given trial, listeners were asked to determine which of two observation intervals contained the signal. If the listener was incorrect, the level of the signal was increased on the next trial. If the listener was correct for two trials in a row, the level of the signal was decreased on the next trial; otherwise, the level of the signal didn't change on the next trial. A block of trials began with the signal presented well above detection threshold. As the listener continues to make correct judgments, the level of the signal is lowered until it comes within the range of the listener's detection threshold. As errors are made, the level increases back into the range of easy detection, and lowers again as correct judgments are made. Following Levitt, we recorded the signal levels at which the direction (increasing or decreasing over trials) reverses. After a fixed number of reversals, the procedure terminates and a threshold is determined by the average of the reversal levels. This measurement is repeated and a final threshold for the experimental condition is reported based on 3-5 blocks of trials.

Spatial audio display

Individualized HRTFs for all subjects were measured using a system based on the HeadZap HRTF measurement system (Begault et al.) using blocked meatus microphones with Sennheiser KE-4 capsules. Measurements were taken in the Spatial Auditory Research Lab at New York University – an acoustically hemi-anechoic room. Measurements were taken using a 1-second sinusoidal sweep sampled at 44.1kHz. For improved SNR, measurements were repeated 3 times at each location, and averaged. The locations measured were every 10 degrees in azimuth, at elevations from -36° to +54° in elevation, at 18° intervals. There were 6 speakers in fixed locations, each positioned at one of the 6 elevations. The subject was located 1 meter from the speakers, seated on a rotating stool. The subject repositioned every time a new location was measured by

rotating on the stool and faced markers on the walls, representing the locations to be measured. Head-Related Impulse Responses (HRIR) with a length of 200 samples were stored. Interaural Time Differences (ITDs) were extracted and stored separately.

Software Environment

Matlab was used to synthesize the infrapitch sources and pink noise maskers, and to render the spatial audio display using the measured HRIRs. It was also used to control the psychophysical procedure.

Experimental rationale

The first experiment was performed primarily to orient the listeners to listening to spatial audio signals and to collect baseline measurements for the detection of infrapitch noise sources in pink-noise maskers. The second experiment assessed the spatial interference on the detection processes when additional sources are present. Both of these experiments presented infrapitch sources and pink noise maskers at known spatial positions. In the third experiment, the locations of the infrapitch source and maskers were unknown.

Subjects

Subjects were recruited from students at New York University enrolled in the Music Technology program. In addition, experiments using classified sonar sounds were performed at the Naval Submarine Medical Research Laboratory using enlisted Navy personnel with prior training in sonar.

Results

Three experiments were performed to investigate the effects of signal, masker, and spatial uncertainty on the ability of listeners to detect a signal in a spatial-audio display. Results of each follow.

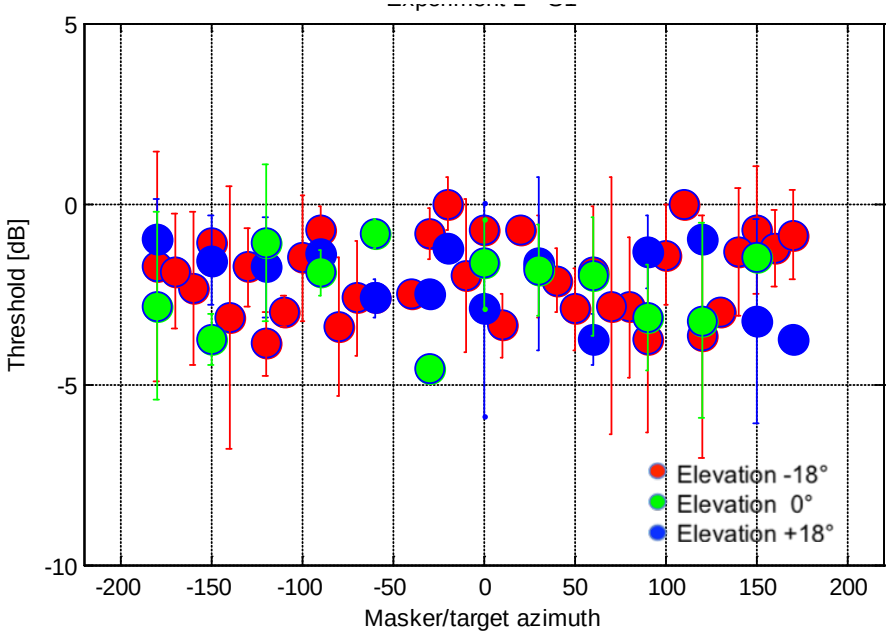

Figure 6. Detection thresholds for one subject at 12 azimuths and 3 elevations.

Experiment One: Figure 6 shows results from one of the NYU subjects for the detection of an infrapitch pink noise target presented at the same bearing as a pink noise masker. Thresholds were measured at 36 different bearings in 10 degree steps. Symbol color in the figure denotes elevation (-18, 0 and +18 degrees). For each elevation, twelve azimuths were measured. Standard deviation bars are shown for each measurement. The present results are typical of those for all NYU and NSMRL subjects: infrapitch thresholds vary over a 5 dB range and are within the expected measurement error. Thus, there is no effect of azimuth or elevation on threshold.

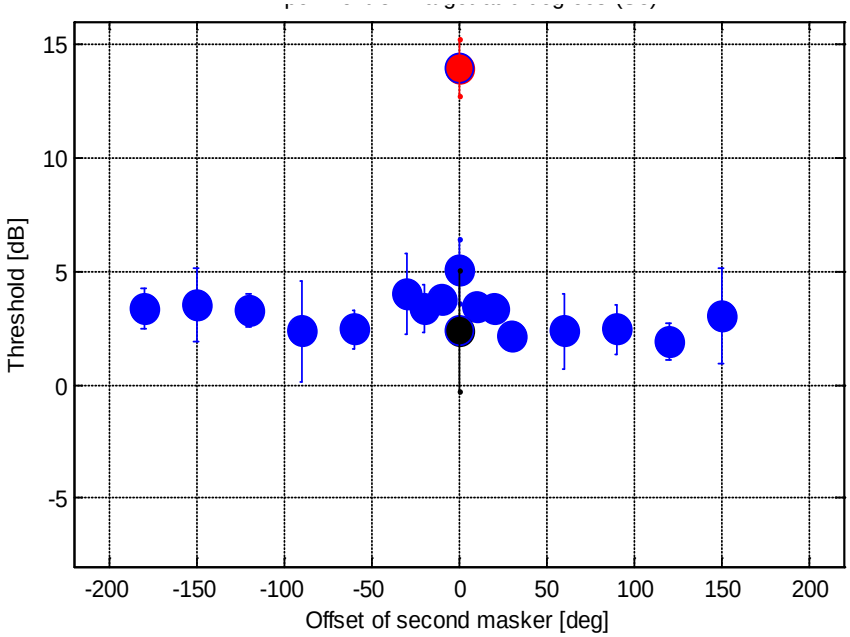

Figure 7. Detection thresholds for one subject with single masker on-bearing (black), at 12 off-bearing locations (blue), and for 12 combined off-bearing maskers (red).

Experiment Two: Having established evidence to support the hypothesis that infrapitch detection mechanisms follow the same expected independence as the more standard spectral-detection mechanisms associated with binaural perception, Experiment Two was designed to assess the degree of interference two or more sources have on the extraction mechanism. A single elevation (0 deg) was used. Thresholds for three conditions were measured. The first condition repeated the threshold for an infrapitch pink noise source in an on-bearing pink noise masker at a bearing of 0 degrees. (In Figure 7 the threshold for this condition is shown in black.) The second condition added to the on-bearing masker a second off-bearing noise masker of equal power. Offsets of this second masker were ±10, ±20, ±30, ±60, ±90, ±120, ±150, 180 degrees. (Thresholds for the second condition are shown in blue.) Finally, the third condition added to the on-bearing masker, 35 off-bearing noise maskers of equal power from -170 to 180 degrees in 10-degree steps. (The threshold for this condition are shown in red.) Threshold is reported relative to the on-bearing masker power.

As shown in Figure 7 for one of the subjects at NYU, the effect of the second masker is to elevate threshold within a neighborhood of target bearing. For the on-bearing case,

when the second masker is introduced with a 0-degree offset, a 3-dB elevation is observed, as would be predicted based on masker addition. The critical bandwidth of interaction appears to be on the order of ±30 to ±60 degrees. For azimuths along the back-side of the head (over ±150 degrees), elevated thresholds are also observed, as would be expected from the acoustics of the cone-of-confusion. Finally, threshold is elevated by approximately 12 dB when there are 36 equal-spaced pink noises.
The results shown are typical of those measured for all NYU and NSMRL subjects with the exception that the signals for the sonar operators at NSMRL were familiar ocean sounds in sea-state noise leading to much lower absolute thresholds of detection in those experiments. In general, the results indicate that the presence of additional sources is disruptive to the periodicity-detection mechanism – the "mainlobe" of the system is fairly wide (60 to 120 degrees of the available 360 degrees and much more consistent with the mathematics of array processing for two sensors). The elevation observed for the 36 equal-spaced pink noise maskers is consistent with this type of "mainlobe" processing. Thus, if any spatial filtering is performed by the binaural system to separate out a signal along a given bearing before undergoing periodicity-detection processing, such filtering is quite limited in resolution.

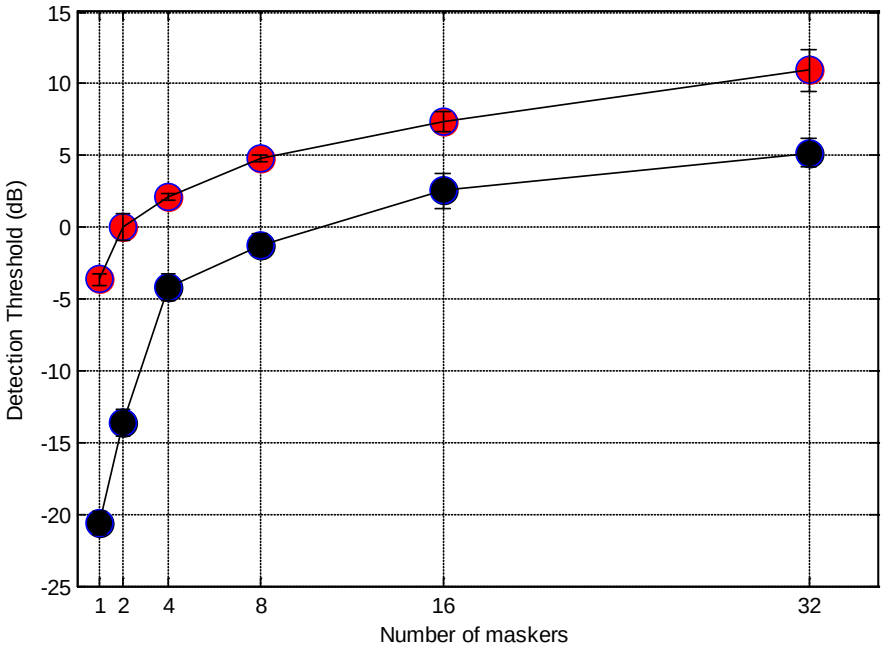

Figure 8. Detection thresholds versus number of maskers for one subject with pink noise sources (black) or other infrapitch sources (red).

Experiment Three. In the final experiment, the locations of the target and the maskers were varied randomly from trial to trial to assess the impact that spatial uncertainty has on the periodicity-detection mechanism. Number of maskers (1, 2, 4, 8, 16, and 32) were varied across blocks. In addition, maskers could either be drawn from pink noise sources (black symbols), as was the case for the first two experiments, or from other infrapitch sources (red symbols). In the latter case, the repetition frequency ranged from 2 to 8 Hz,

excluding the 5-Hz repetition frequency used for the infrapitch signal. Signal and masker locations were rendered at azimuths that were drawn randomly without replacement over -170 to 180 degrees in 10-degree steps with a fixed elevation of 0 degrees.

Consistent with the mainlobe characteristic of the periodicity-detection mechanism that was observed in Experiment Two, the masking curves for both masker types exhibit considerable saturation as a function of the number of maskers. The effect of having "decoys", that is, maskers that engage the periodicity-detection mechanisms in the same manner as the infrapitch signal, appears largest for the case of 1 or 2 maskers, but then shows approximately a 5-dB penalty regardless of the number of maskers. When comparing the overall magnitude of masking shown, there is little evidence that not knowing the location of the maskers and target has any effect on performance. To model this precisely would require further research that is not within the tasks we were assigned to accomplish.

These results show that:

1) Listening for periodically repeating signals (2-8 Hz rates) whose spectral or temporal features are otherwise unknown, follows the same pattern of binaural masking as highly-certain signals.
2) Prior knowledge of the spatial configuration of the signal and maskers does not appear to be necessary to detect the periodically repeating signal.

Results of the three experiments indicate that the auditory processes involved in extracting and recognizing ocean-born rotating machinery noise are not disrupted under spatial audio display. In addition, when comparing these results to those obtained for multi-talker speech tasks, the results suggest that spatial awareness is far more robust to the effects of multiple sources and spatial uncertainty than is speech communication. Both of these results support the use of spatial audio displays for maintaining situational awareness of high-complex, multi-source environments.

Discussion

As listeners, we take our perception of sounds in space for granted. Our folk knowledge tells us that we hear sounds coming from any direction whether or not we are looking in that direction at the time the sound occurs. Folk knowledge also tells us that this cross-sensory modality independence holds within the auditory modality as well: while attending to a sound in front of us, we are not prevented from hearing and responding to a warning sound, for example, to our left or right. The number of sources we can monitor simultaneously would appear to be arbitrarily large. We readily perceive an audience applauding, a crowd cheering, or an urban street scene rich in traffic, talking, and a variety of transients. When properly trained, orchestral conductors demonstrate the basic perceptual capacity of our auditory systems in their ability to monitor 100 instruments or more. Audio engineers mix multiple-track recordings of ensembles of performers and listeners acknowledge the improvements in hearing the nuances of each instrument when going from mono to stereo to surround sound.

In light of our folk knowledge and the working examples above, it is surprising that contemporary experiments in the design of spatial audio displays suggest our perceptual capacities are very limited. For the past decade, Brungart and his colleagues have studied the use of spatial audio displays for improved speech communication in multi-talker environments (Brungart, 2001; Brungart and Simpson, 2002; Brungart et al., 2009). They have demonstrated that listeners are so limited in their spatial abilities that they cannot monitor more than 5-7 locations simultaneously.

One way to reconcile our common sense notions of spatial hearing with Brungart's findings is that both are tapping into spatial hearing at different levels. At the most basic level, we propose that spatial hearing supports spatial cognition or awareness of our environment. Once recognized as a source located at a particular point in space, spatial hearing supports the processes involved in extracting properties about that source. Thus, the "where" of sources taps into the most basic level of spatial hearing whereas the "what" of sources taps into the extraction of information from sources in the environment. All of the "folk knowledge" examples above focus primarily on questions of "where"; questions of "what" are either categorical (e.g., is it a cheer, a hand clap, a car horn) or tap into very primitive perceptual properties of sound (e.g., pitch, duration, loudness). In contrast, Brungart's experiments are not assessing "where" as much as they are assessing very detailed questions of "what": the task is to extract pairs of words spoken by one speaker in the presence of one or more other speakers. Indeed, Santoro and Wakefield (2005) reported very different results to those of Brungart by asking sonar listeners to determine whether a given target source was present in an ensemble of distractor sources. Sources, in this case, were surface ship sounds rather than talkers, and the information to be extracted was "ship identity" rather than words. Santoro and Wakefield did not observe any performance degradation for as many as 10 sources. Evidently, ship identity is more robust to the presence of multiple sources than speech recognition.

Operational Tasks

The greatest concern we had during the first year of this study was whether the tasks for which spatial audio was to be employed in sonar operations were closer to the most primitive level of spatial perception (the "where") than the most detailed level of source perception (e.g., what word was spoken). In our discussions with members of the operational research team at APL, we identified several relevant features of "typical" listening situations:

- the acoustic field, in any given direction, will be dominated by sea-state noise;
- operators will be monitoring their environment for the emergence of sources that, to a first approximation, may only be identified by rotational machinery noise rather than highly-refined spectral-temporal features;
- emergent sources may come from any direction;
- there may be multiple sources of interest, and these may also come from any direction;

- the number of directions of interest at any given time is arbitrarily large.

Among all the features, the one that caused us greatest concern was that the spatial audio display would be required to present a large number of sources. Given the findings of Brungart, it was very important to determine whether the extraction of source features (rotational machinery noise) was robust to noise and other machinery-noise sources for large numbers of sources.

Conclusion

Spatial audio displays can be used to support a number of different tasks. There appear to be two different conclusions about the use of spatial audio for "recognition" experiments: in speech communications, spatial audio appears to have limited utility, whereas in the less demanding context of recognizing a target surface ship in the presence of others, spatial audio substantially improved performance over single-channel audio. The question is whether operational tasks (find the target ship) are closer to those of speech communications (in which case, our research would need to explore methods for source reduction in the spatial display) or spatial awareness (in which case, our research would need to address the limitations imposed by auditory masking on performance).

We chose to work with Infrapitch noise sources to capture the hardest problem an operator might face: listening for the presence of sounds from rotating machinery somewhere in the environment in the presence of other such sounds. The results presented in the three figures of this section, along with many more from our monthly progress reports and debriefs, allowed us to conclude that the operator tasks are likely to benefit from spatial audio display of very large numbers of sources, in contrast to what has been demonstrated for speech communication where the number of sources is severely limited.

We note that 1) the present results have bearing on the ongoing debate within the psychoacoustics community concerning the processes involved in segregating acoustic information into auditory streams and 2) the often-cited reason for not pursuing "large-source-count" spatial audio displays is very task specific. With regards to the first point, our data have direct bearing on those processes that are likely responsible for "source build-up" phenomena in auditory streaming. Results from our experiments involving spatially-located infrapitch sources should provide strong tests of some theories for auditory stream segregation. Along more local lines, the data are worthy of further detection-theoretic modeling and could prove interesting to model, in particular, using Durlach's Equalization-Cancellation approach.

With regards to the second point, the most systematic Human-Computer Interaction (HCI) studies of spatial audio displays are those utilizing the paradigm originally set forward by Brungart and his colleagues in 2001. These studies concentrated on command center tasks by addressing the use of spatial audio in maintaining several lines of simultaneous speech communication. What is interesting about these studies is the tertiary role that spatial position plays in the tasks. Location is important only in so far as

it supports the listener's ability to process speech at one location to the exclusion of all others. This is a very different use of "space" when compared with the much broader and substantially older literature in spatial cognition or cognitive maps. More than 50 years of research in such disparate disciplines as experimental psychology, geography, cognitive science, neuroscience, artificial intelligence, environmental psychology and robotics have demonstrated that "location" is the necessary stimulus feature that supports a wide variety of behavioral tasks, including many that have been loosely described as examples of "situational awareness".

Thus, the appropriate HCI studies for spatial audio displays shouldn't be limited to those that use "location" only for the purposes of improving speech communication. As our work has demonstrated, it is no wonder that Brungart's research shows so little advantage in a spatial audio display since the speech processing task is, itself, easily disrupted and the functioning of the binaural system is inherently limited to the mathematics of "two sensors", making disruption all the more likely. Rather, the more appropriate HCI studies for spatial audio must investigate the spatial dimension, much as has been the case for all other research studies in spatial cognition. To this end, the orientation modules we have developed are strongly oriented towards questions of spatial processing and are consistent with many of the possible tasks demanded by robust "situational awareness".

NAVIGATION IN A VIRTUAL AUDITORY ENVIRONMENT

Methods

An experiment was designed to assess the extent to which auditory search in a virtual acoustic environment (VAE) can be mediated through an avatar interface. The VAE was comprised of acoustic sources arranged along a circle in an otherwise anechoic environment. Participants could move and orient through this environment either directly, by walking and turning their head, or indirectly, by moving the location and angular orientation of an avatar on a computer display presented in a top-down perspective. In what follows, the former will be called *natural mediation* and the latter will be called *avatar mediation* of user position in the VAE.

The task required that participants locate a source in the VAE by moving to the location of that source. To acclimate participants to the apparatus, the experiment was conducted in two phases. During the training phase, a single source was presented during a trial and the participant moved from the center of the circle to the location of the source as quickly as possible. During the test phase, four sources were presented during a trial and the participant was to move to the location of each source until all four sources were found. Because it draws upon the standard means by which we, as listeners, navigate through our environment, we hypothesize that natural-mediated search will require fewer trials than avatar-mediated search to reach asymptote for the training phase. Nevertheless, because both forms of mediation engage a common representation of auditory space, we expect that the asymptotic search strategies of each will be similar.

When multiple sources are present, it was not clear how search times should be affected. An increase in the time it takes to locate the first source would be expected if the presence of multiple sources interferes with the cues used to locate any one source. Alternatively, a participant may choose to minimize total search time by using a portion of their first search to establish a general mapping of all the sources before moving to the first source. In the absence of interference or a global strategy, the time it takes to locate the first source during the test phase should be the same as the asymptote reached during the training phase.

Finally, we were interested in whether some users are generally faster than others when performing an auditory search and in the strategies they use. Accordingly, each participant was tested under both forms of mediation. Half the subjects were trained and tested first under natural mediation, before going on to training and testing under avatar mediation, while the other half underwent initial training and testing under avatar mediation. We hypothesize that experience in either modality (natural or avatar mediation) will transfer to the other as evidenced in fewer trials to reach asymptote when shifting to the alternative modality and that there will be a high degree of correlation between fastest and slowest performers across modality.

For both training and test phases of the experiment, a trial began with a source (or sources) positioned randomly along a fixed circle placed horizontally in the 0-degree elevation plane and the participant positioned in the center of that circle. Participants were notified by a diotic auditory cue when they arrived within a fixed radius of the source. During the training phase, a single source was presented and the participant re-centered him or herself after notification to begin another trial. Training continued until a participant's current and past four search times had a standard deviation of 2.5 seconds or less.

The test phase consisted of four sources. At the beginning of a trial, participants were informed which source they should search for first by a diotically-presented four-second sample of the selected source. Following the cue, the four sources were presented and the participant began their search. Upon successfully locating the first source, the sources were turned off, and the second cue was presented, after which the sources were turned back on again. This sequence continued until all four sources were located.
When a participant finished both the test and training phases for one modality, they repeated the procedures for the alternate modality. The testing was conducted at New York University. Eighteen paid volunteers participated in the experiment. Half began with training and testing using the natural mediation while the other half trained and tested on avatar mediation first. Training and testing under both modalities took approximately 75 minutes. Each participant completed the experiment in one session.

Results

Subjective testing was conducted to study whether we can use an alternative interface to the head tracker during a search task in an auditory environment. A traditional 6DOF tracker was compared to a mouse/keyboard interface in a navigation and search task where subjects were asked to find acoustic sources. Results show that:

1) The number of trials necessary to reach optimal performance with each interface was similar, however, there is clear evidence of learning to use the mouse to interact with the acoustic environment.
2) The optimal search times reached using both interfaces was very similar
3) Regardless of whether the search was within a one- or four-source context, the search times for the first source are very similar.
4) There is no significant difference in the quality of a user's search strategy using the head tracker as compared to the mouse interface.
5) The number of sources in the environment does not have an impact on the search time to find the target.

Results of this experiment indicate that an alternate interface can be used without sacrificing the target search time in an auditory environment.

There are alternative interfaces that can be used as a means of interaction and navigation through auditory environments. In this paper we discuss results of a study that compares the use of a traditional 6DOF head tracker to an avatar interface, through the use of a mouse and keyboard as a navigation medium. We compare human performance in a search task to find a source (or sources) within a single and multi source context auditory environment during a subjective experiment.

Analysis was performed on the data from the training period, the first source within the one-source and four-source context, and all sources within the multi source context. Typical results for the training period are shown in Figure 9 for the mouse interface (a) and the tracker interface (b). The white bars indicate the trial at which optimal performance was reached. The results show evidence of learning before reaching optimal performance use the mouse interface, but little improvement in performance over time while using the tracker interface.

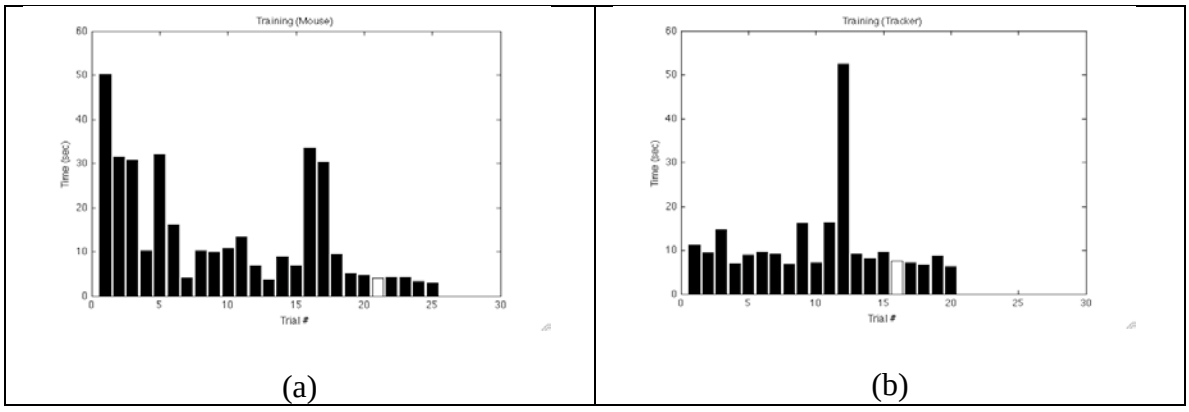

Figure 9. Example of training time results for a subject using the avatar mediation (a) and natural mediation (b). The subject was presented with the natural mediation first. The trial marked in white represents when subject has reached optimal performance.

The number of trials it took subjects to reach optimal performance varied. The scatter plot in Figure 10 shows the trial number at which performance was reached. The mouse

interface first subjects are represented using ‘x’, the head tracker interface first subjects are shown in ‘o’.

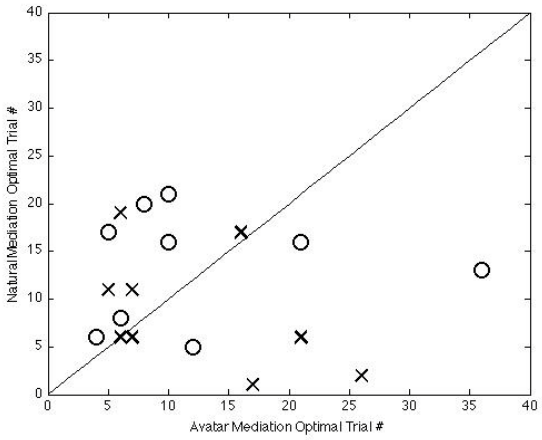

Figure 10. Optimal trial number for the avatar-first (x-marker) and natural mediation-first (circles).

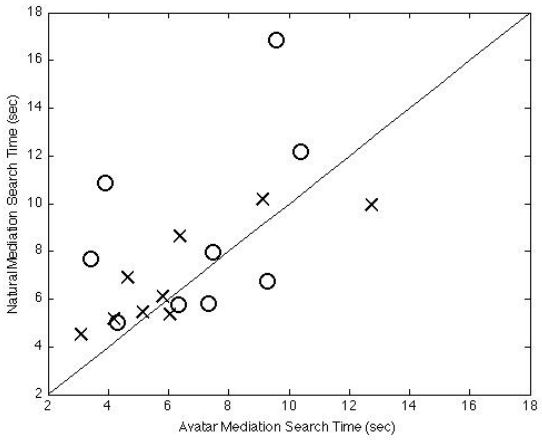

Figure 11. Single-source environment search times for avatar mediation-first (x-marker) and natural mediation-first (circles).

Analysis of the search times within the single-source context, reveals that subjects tended to spend an equal amount of time finding a single source regardless of whether they used the mouse or tracker interface. These results are summarized in Figure 11, for subjects using the mouse interface first ('x') and head tracker first ('o').

When comparing the search time results between the training session and the test trials in the single-source environment, we see similar performance between the two phases of the experiment. Figure 12 compares training search times (x-markers) and the test search times (open circles) for the two interfaces. In most cases, very similar results can be seen during the test trials, as when optimal performance is reached during training. In other words, it appears that once a subject reached a certain level of performance during the training phase, they managed to maintain this level.

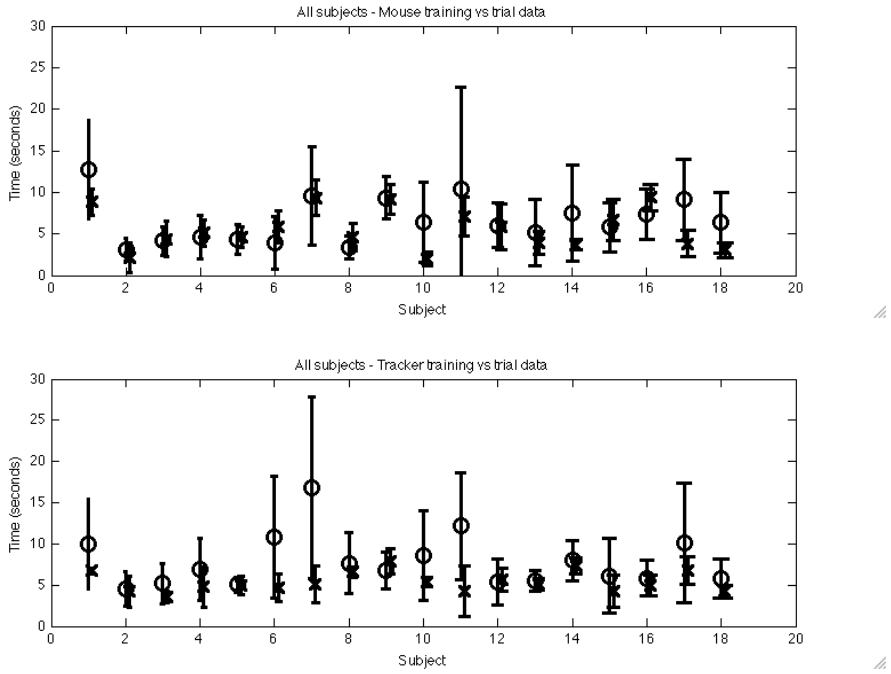

Figure 12. Error bar plot comparing results of training data (x-markers) to test data (open circles) for the avatar (upper) and natural (lower) mediation.

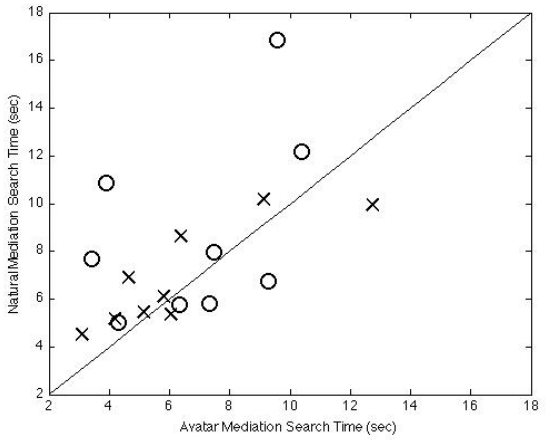

Figure 13 Single-source environment search times for avatar mediation-first (x-marker) and natural mediation-first (circles).

The mean search times during the testing phase for the single-source environment are presented in the scatter plot in Figure 13, for the avatar (x-markers) and natural (open circles) mediation. For many subjects (almost 50%) the search times for both types of mediation for each subject are very similar. A subject tended to spend an equal amount of time finding a single source regardless of whether they used avatar or natural mediation. This is consistent with our results from the training sessions, where we saw a similar search time for both types of mediation. However, when looking at the raw data for all subjects for the 1-source context, we see an overall increase in search time going from the avatar to the natural mediation. When looking at all subjects, the mean search time for the avatar is 6.2 sec, and 7.8 sec for the natural mediation. Subjects who were presented with the avatar first, show a mean response time of 6.4 sec with the avatar mediation, and 6.9 sec with the natural mediation. Subjects who were presented with the natural mediation first have a response time of 6.9 sec using the avatar, and 8.8 sec using the tracker.

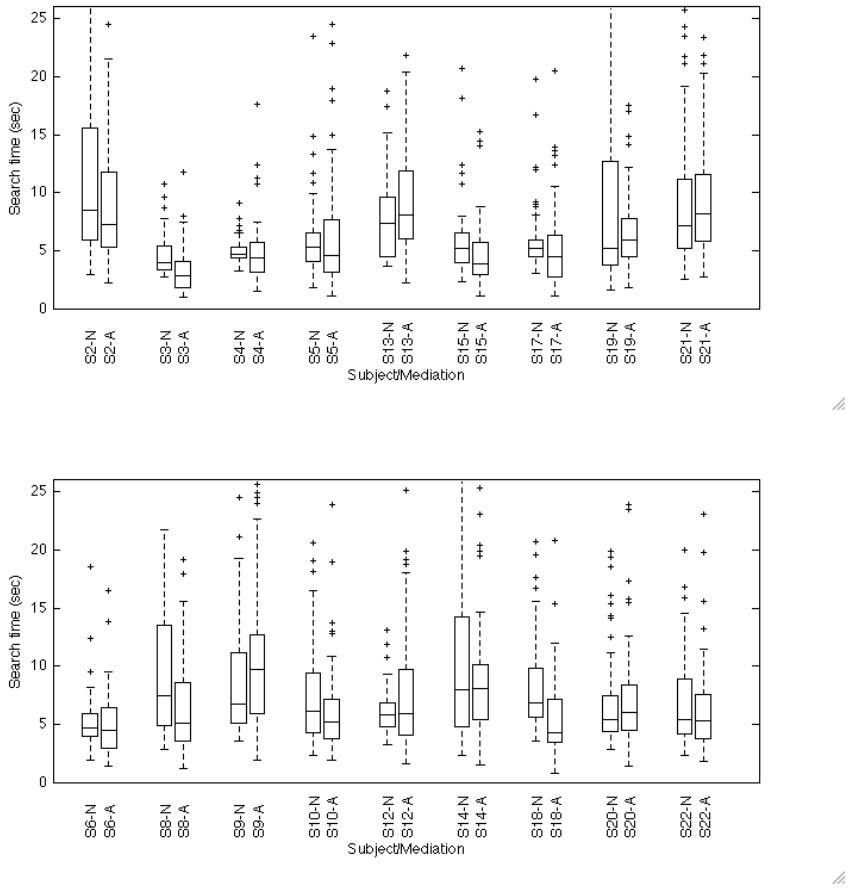

Figure 14. Boxplots of search times for each subject for the natural mediation (marked as N) and avatar mediation (_A) for subjects who were presented with the avatar mediation first (upper), and those who were presented with the natural mediation first (lower).

In the multi-source context, we see similar trends in the search times of the avatar and natural mediations. Subjects show similar search times for both mediations, regardless of which mediation was presented to the subject first. In the multi source context, we see an elevated range of responses. This could be caused by the fact that there were four times as many responses made by subjects.

Results from the analysis of the search time for each source, in the single source and multi source context, are presented in Figure 15. The x-axis on the figure indicates the source number, the mediation, and the context in the format "Src#_Mediation_Context"). On the left, we see the pair results of the natural (Src_N_S) and avatar (Src_A_S) search times in the single source context. Further to the right of the graph we see the result pairs for the first through the fourth source search times in the multi source context (Src1_N_M through Src4_A_M). The median search times for each mediation and context are listed in Table 1. We see here that the search times are nearly identical regardless of the context or the source number in the search, indicating that the presence of additional sources does not pose an advantage nor a disadvantage to the search.

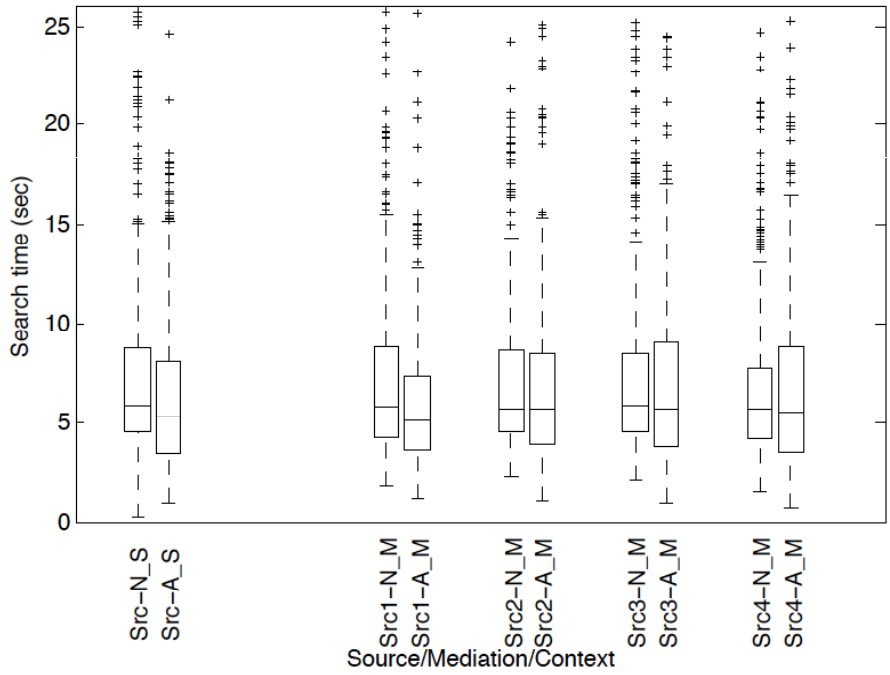

Figure 15. Comparison between the single-source context (_S, on left of graph) and the multi-source context (_M, on right side of graph), for all 4 sources. Results are presented for all subjects.

Table 1. Median search times for the natural and avatar mediations, for the single and multi source contexts.

	Natural (4-src)	Natural (1-src)	Avatar (4-src)	Avatar (1-src)
All sources	5.71	5.88	5.42	5.28
Source 1	5.74	5.88	5.15	5.28
Source 2	5.70		5.65	
Source 3	5.86		5.72	
Source 4	5.66		5.50	

Search strategies

To further evaluate the subjects' performance using each form of mediation, we analyzed the paths taken by each subject when finding the source. These paths are indicators of the search strategies used by the subject, and give us insight into how well the user's spatial knowledge of the interface is being utilized. Prior research (Buechner et. al, Hill et. al, and Thinus-Blanc & Gaunet) have classified spatial search patterns into those that indicate novice search performance and those which indicate a more experienced search technique. It is by these classification schemes that we have categorized each subject's path data. Figure 16 shows an example from our data of a path that would be classified as a novice (left) strategy and a path that would be classified as an experienced (right) strategy.

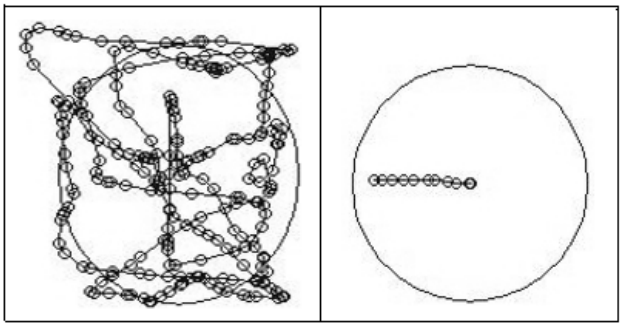

Figure 16. Classification of search strategies. The path on the left is classified as a directed random strategy and the path on the right is classified as an enfilading strategy.

Discussion

The degree of interaction between the participant and the sensorimotor level is a key factor in the success of immersing the participant in the environment. In no other application is interaction more important that the simulation of spatial auditory environments, where the relative location of the participant and the sound sources is a key factor of maintaining a stable and realistic image, and where interaction can dramatically improve the quality of the spatial auditory image. In auditory environments, interaction has traditionally been accomplished with the use of a head tracker. The orientation information (and position information, if available) obtained from the tracker is used to simulate sounds in the virtual auditory environment. However, there are several disadvantages associated with the use of head trackers, including cost, sensitivity to calibration, specialized software development, and their sensitivity to the environment in which they are used. For example, many trackers use disturbances in electromagnetic fields to determine position and orientation information, or use magnetic north as a source of calibration – these trackers become practically unusable in environments with high metallic content, such as submarines.

Results from the training phase, during which subjects familiarized themselves with each interface and the task, show that in many cases, the number of trials necessary to reach optimal performance with each interface was similar. However, looking closely at the trial data we notice that there is clear evidence of learning to use an avatar to interact with the acoustic environment. Such steep learning curves were not seen in most subjects under natural mediation. These results are independent of whether the avatar mediation was presented to the subject as the first or the second interface and suggest no transfer of experience across the two interfaces.

The asymptotic search times achieved during training were very similar in both interfaces. During the testing phase we saw an increase in the search times in both the single and multi source context conditions (from 5.29 sec to 6.2 sec with the avatar, and from 5.11 sec to 7.8 sec). Although the training and test search times in most subjects were very similar, in a few subjects we observed an increase in search times in the testing phase of the experiment, which could be due to fatigue. Further testing is needed to validate the cause of the search time increase.

Regardless of the number of sources in the context (one or four sources), results show nearly identical search times. This suggests that the number of sources in the background does not create a distraction for subjects, nor does it help them by allowing them to build a mental map of the complete environment.

Congruent with the search time data, the search strategy data also indicate that there is no clear difference in the quality of a user's search strategy under natural mediation compared to avatar mediation for finding a single source in a four source-environment. Small differences exist in the proportion of experienced search strategies used, while training as well as in the one source environment. Although these differences can be teased out, they are not significant enough to suggest that one form of mediation is significantly superior to another.

The experiment was setup in a room where the physical configuration and the limitations of the sensitivity and range of the tracker used in the natural mediation limited the physical space during testing to a radius of 1.5 meters. Although we have not performed any testing in different sized configurations, we speculate that the size of the effective area during the testing was one of the contributing factors to the similar time scales of the results. Had the area been much larger, the physical constraints of human movement would have most likely produced different results, as it is doubtful, for example, that a virtual acoustic source placed somewhere in a football field would be found by most players in under 10 seconds! The key finding of our experiment is that the only penalty in using an avatar to explore one's acoustic environment is that of learning to use the interface in the first place. Once learned, participants appear to use it as effectively as they would their own bodies in exploring a new acoustic space.

In summary, the following conclusions can be drawn from the experiment data and analysis:

1) The number of trials necessary to reach optimal performance with each interface was similar, however, there is clear evidence of learning to use the mouse to interact with the acoustic environment.
2) The optimal search times reached using both interfaces was very similar
3) Regardless of whether the search was within a one- or four-source context, the search times for the first source are very similar
4) There is no significant difference in the quality of a user's search strategy using the head tracker as compared to the mouse interface
5) The number of sources in the environment does not have an impact on the search time to find the source.

A full paper was presented at the 16th International Conference on Auditory Displays on June 11th, 2010 (Roginska et al., 2010d). A second paper focusing on the results of the analysis of the multi source context was presented at the UHSI symposium July 27-29th in Providence, RI (Roginska et al., 2010b).

Topics for further investigation

Our research has shown that listeners can replace the natural head-motion feedback they rely on to resolve ambiguities in their auditory environment with mouse-driven feedback. This finding,

however, involved both head rotation and body movement. We are concerned that key experiments in which translational motion in the environment is eliminated have yet to be done. Without the results from these experiments, it is not clear that anything but head-trackers will provide the necessary feedback for the user. Should a system be deployed which replaces a head-tracker with some alternative means of rotating the head through the environment, it is unclear what types of operator error will result. We strongly recommend a research and development program that explores a variety of physical devices for rotating the user's orientation within their virtual auditory environment and compares performance on realistic tasks with that achieved with the standard, more costly and less robust, head-tracking systems.

HRTF MEASUREMENTS AND SELECTION

Methods

The quality of the measurement of a head-related transfer function (HRTF) is one of the key factors in presenting a good quality spatial audio simulation to a listener. Currently, the commonly used method to acquire HRTFs is through acoustic measurement. The HRTF measurement system at NSMRL consists of three custom-built components: a speaker array for stimulus presentation, a microphone/earplug assembly designed for signal acquisition in the ear canal, and a suite of MATLAB software that implements a system identification procedure using Golay codes. In production, the listener's HRTFs can be determined at 270 different spatial positions within 45 minutes.

The speaker array consists of an array of 15 two-way speakers mounted on an arc suspended vertically in the 10m x 10m x 10m anechoic chamber at NSMRL. The speakers are located at 18-degree intervals along the arc, yielding samples in elevation from -36º to 90º in front of the listener and then back down to -36º behind the listener (apart from some measurement time advantage, the extra three elevation data points measured by this apparatus are the main difference between it and the 6 speaker array used at NYU). The listener sits with their head positioned in the center of the arc which coincides in all three dimensions with the center of the anechoic space. The arc is constructed from PVC tubing buttressed by a solid steel pipe. The PVC tubing is filled with sand and is covered with acoustic diffusing foam to reduce unwanted acoustic reflections and vibrations during the measurement process. The arc is circular, with an approximate radius of 76.375 inches, and the center of the arc is approximately 73 inches above the chamber's wire-mesh flooring. To minimize reflections, the speakers are covered in acoustic diffusing foam.

Different azimuths are sampled by rotating the listener with respect to the arc. To accomplish this, the listener is seated in a workbench-height rotating chair, which is covered with acoustic diffusing foam to reduce reflections and vibrations. To aid the alignment of the subject's head to various elevations and azimuths, two methods are used. First, a Polhemus head-tracker is placed on the subject's head during measurement. Readings from the head-tracker are used to verify the specified azimuths and to check head alignment in elevation. Second, visual markers are placed along the chamber walls every 10 degrees. Using these visual markers, listeners were able to rotate themselves to within 1-2 degrees of the desired (tracker monitored) azimuth.

The signal acquisition system uses the Knowles FG3329 microphone, a small, cylindrical microphone intended for use in in-ear hearing appliances. This microphone is glued into a modified over-the-counter silicone earplug, designed to fit children's ears. This earplug/microphone assembly fits comfortably and completely into the ears of subjects to provide for "blocked ear canal" measurements. The microphone signals are conditioned by a custom-built pre-amplifier, and then digitally sampled (16 bit resolution) at a 50 kHz sampling rate by Tucker-Davis Technologies System II hardware.

MATLAB scripts implement system identification using Golay-codes. Ten unique, 512-point Golay code pairs with an amplitude of ±30000/32768 are used to measure the left and right HRTFs at a single spatial location. To reduce the effects of uncontrolled head and/or subject rotation during measurement, the two Golay codewords in each Golay pair are presented consecutively, where each codeword is preceded and followed by 512 samples of silence to allow for the decay of reflections. Signal to noise ratio at each spatial location is maximized by automatically adjusting the gain of the amplifier attached to the presentation speaker so that the signal at the microphone had a prescribed amount of energy.

This measurement process is time-consuming and the requirement for the listener to remain seated for an hour (or more) can introduce errors due to noise, fatigue, listener motion, or other errors associated with human use. The HRTF process is also costly, requiring specialized equipment and facilities. Examples such as incorrect microphone placement or lack of system calibration can lead to severe measurement inaccuracies and alter measured HRTFs, thus leading to distorted auditory images and unusable HRTFs.

Hence, a method was developed to analyze a measured HRTF data set, diagnose the data set, and report any locations which may have potential measurement errors. The tool *VerifyHRTF* was developed to serve as a utility with a GUI to view, diagnose, and correct a measured HRTF dataset. The following figure is a screenshot of *VerifyHRTF*.

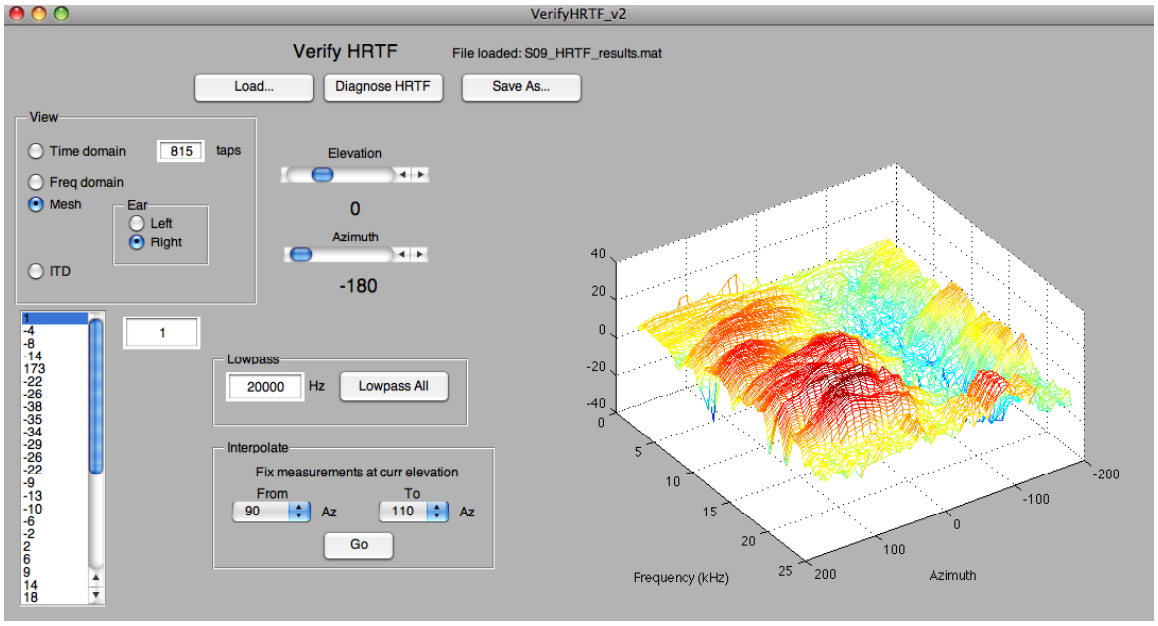

Figure 17. Screenshot of the VerifyHRTF tool

An HRTF data set can be loaded to the utility. There are navigation sliders to navigate through the different azimuth and elevation locations.

There are four main view modes:

- time domain, with selectable filter length;
- frequency domain (magnitude);
- mesh: a 3-dimensional plot that represents the magnitude of all azimuths at a selected elevation. The left or the right ear can be selected for plotting
- Interaural Time Difference: plot of all ITDs at a selected elevation.

There are three main correction tools in this utility. The first allows to manually change/correct the ITD. Numerical ITD values are listed in a list (shown on the left side of the screen). Each value can be selected and manually edited. The visual display immediately reflects this change so that results can be viewed.

One of the issues observed with some HRTF measurements is the presence of abnormal high frequency data. When the problem occurs, this can be easily seen in either the “Frequency domain” or “Mesh” viewing modes. A tool has been developed to lowpass all HRTFs, with a specific cutoff frequency.

The third correction tool allows the user to interpolate data between two measured locations. The user selects two locations containing correctly measured data and the location points between these are calculated.

As a first step, we have defined a method to analyze HRTF sets and look at the ITD information. We are comparing the measured ITDs with predicted ITDs based on a spherical head model. Using the spherical head model, we can find a best fit for the ITD and use this as a parameter to determine the head size of the subject as well as other characteristics. A sample plot of the predicted ITDs based on a head radius of 8.9cm and measured ITDs for a subject is presented in Figure 18.

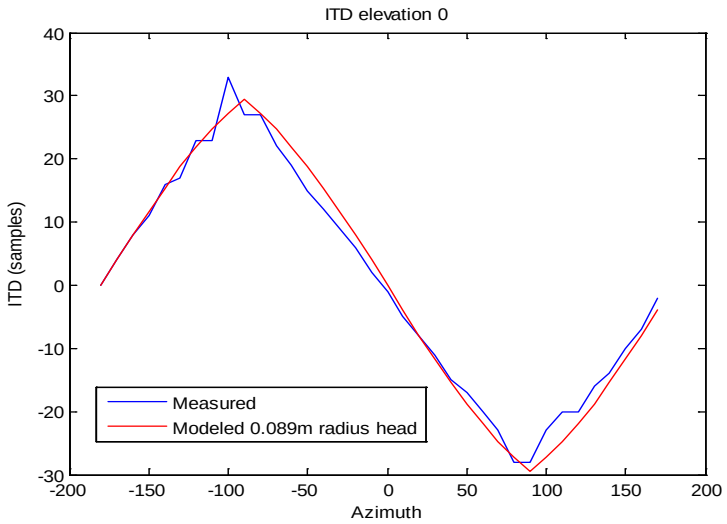

Figure 18. Plot of the comparison of the measured (blue) and modeled (red) ITD.

Using a similar approach, we have also developed a method to analyze a measured HRTF data set, determine the "goodness" of the measurement, and report any locations which may have potential measurement errors. Figure 19 shows an example of an HRTF dataset with possible measurement errors. When compared to the predicted ITDs, it can clearly be seen the location with possible measurement error. After running our current script, the following report is obtained at this elevation:

>> Possible HRTF measurement error at Elevation: 36 Azimuth 70

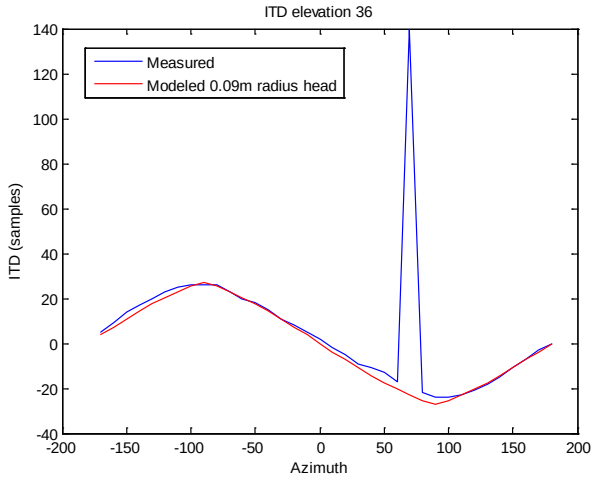

Figure 19. Example HRTF with measurement errors.

A further look at the spectral plot (below) reveals more details about the problem.

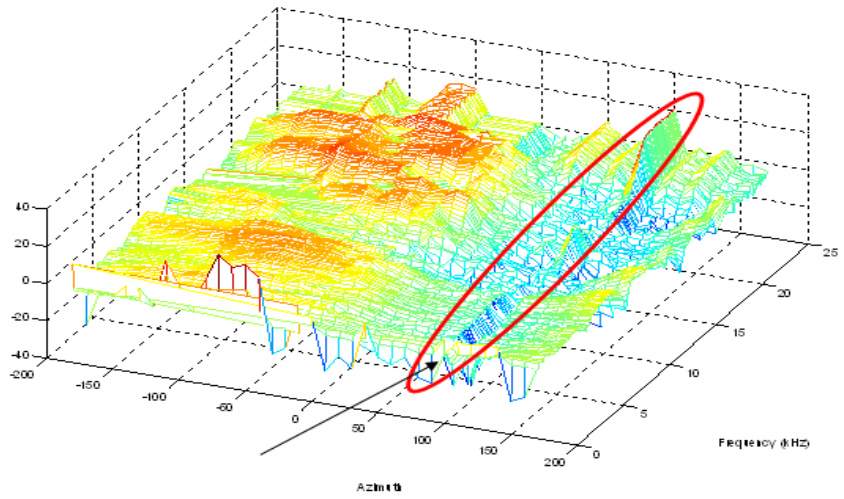

Figure 20. HRTF spectral plot.

The Verify HRTF utility is very useful in examining the objective quality of a measured HRTF and can quickly assess whether there were problems during the measurement procedure. The additional tools developed can also fix some (not all) of these issues or enhance the measurement. The basic question still remains whether individually measured HRTFs are necessary for optimal perception of spatial audio or whether there are alternative HRTFs that we can use without compromising performance.

HRTF Selection Technique

Due to the uniqueness of the size and shape of human bodies, heads, torsos and pinnae, everyone has a fingerprint-like set of HRTFs. The accuracy of HRTFs directly impacts the quality of the simulated spatial sound. Most accurate HRTFs are acoustically measured on an individual. This process can be time consuming, requiring a person to stay seated for an hour or longer, which can lead to fatigue. Measurements can also introduce noise and error due to operator or equipment malfunction, leading to unusable measurements. In addition, the process is costly, requiring specialized equipment and acoustically treated facilities.

Generic HRTFs measured on a mannequin such as KEMAR have been used, but have resulted in poor externalization, degraded localization accuracy and an increased amount of front/back and elevation confusions.

Eliminating the need for individualized HRTFs, while not compromising the quality of the spatial auditory image, would lead to the greater accessibility and better overall experience of spatial audio. In no other applications is this truer than in mission-critical applications.
In a two-phase study conducted during this reporting year, we investigated whether there exist common HRTFs listeners choose when they are asked to select among a database of HRTFs. Second, we investigated whether listeners select their individualized HRTF when this

individualized HRTF is presented as one of many datasets. Third, we looked at whether the type of signal listeners are listening to has an influence of which HRTF datasets are selected.

Phase I

A database of 27 HRTF datasets was collected from publically available databases. Additionally, the subject's individual HRTF was added to the collection of HRTFs, for a total of 28 datasets. The experiment was divided into two parts: HRTF measurement and listening test. HRTFs were measured using a system based on the HeadZap measurement system. HRTFs measured on a subject during the first part of the experiment were included in the general database of HRTFs. The task during the listening test asked the subjects to select all the HRTFs that fulfilled the criterion they were listening for. The listening test was divided into three stages. At each stage a different criterion was the focus of the listening task. Subjects were asked specifically to pay attention to the criterion and select all the intervals where the criterion was fulfilled. The criteria included:

- Externalization: source appeared to be coming from outside the head.
- Elevation discrimination: capability to discriminate a source at a high and low elevation, at a fixed azimuth.
- Front/back discrimination: capability to discriminate a source in front from a source in the back, along the cone of confusion.

The goal of the listening test was not for subjects to make judgments about the absolute location of the stimulus. Rather, the aim was to judge the relative perception of the location of the signal to a reference signal. The sequence of the stimuli changed during the 3 stages of the listening test. During the first stage of the test, where the discrimination criterion was externalization, the reference signal was a monophonic infrapitch signal processed with the 0° azimuth, 0° elevation HRTF (in order to eliminate any difference in signal coloration) with channels cross-summed. Subjects first heard the reference signal followed by a series of 5 signals spatialized at randomly selected locations on the horizontal plane from the following azimuths: ±150°, ±120°, ±90°, ±60°, ±30°. The sequence of locations was the same for all intervals within a trial.

During the second "elevation discrimination" stage, subjects heard 5 pairs of stimuli. Each pair included the stimulus processed at a random azimuth (selected from the same azimuths as listed above), at +36° and -36° in elevation. Subjects were asked to discriminate between the high and low elevation signals and select all intervals where they perceived a difference in elevation. High and low elevations were presented randomly. The azimuths for all intervals were kept consistent through a single trial.

Similarly to the second stage of the test, the third stage consisted of 5 pairs of stimuli. All signals were processed on the horizontal plane (0° elevation). Each pair included two signals along the cone of confusion, one in front the other in back. The locations were selected from the following azimuths: ±150°, ±120°, ±60°, ±30°. Subjects were asked to select all intervals where they could discriminate front from back. The front/back locations were chosen randomly. The locations were kept across all intervals in a single trial.

An HRTF was presented 3 times at each stage of the test. Only HRTFs that were selected at least 2 out of the 3 times during a stage were passed on to the next stage of the listening test.

Apparatus Individualized HRTFs for all subjects were measured using a system based on the HeadZap HRTF measurement system using blocked meatus microphones with Sennheiser KE-4 capsules. Measurements were taken in the Spatial Auditory Research Lab at New York University – an acoustically hemi-anechoic room. Measurements were taken using a 1-second sinusoidal sweep sampled at 44.1kHz. For improved SNR, measurements were repeated 3 times at each location, and averaged. The locations measured were every 10 degrees in azimuth, at elevations from -36° to +54° in elevation, at 18° intervals. There were 6 speakers in fixed locations, each positioned at one of the 6 elevations. The subject was located 1 meter from the speakers, seated on a rotating stool. The subject repositioned every time a new location was measured by rotating on the stool. Head-Related Impulse Responses (HRIR) with a length of 200 samples were stored. ITDs were extracted and stored separately.

A graphical user interface was designed and developed in Matlab. The interface was used to setup the experiment, as well as for user interaction during the listening test (see screenshot in Figure 21). The experimenter loaded the HRTF database and selected the individualized HRTF from which the ITD was extracted and used for all HRTF datasets during the listening test. The stimulus used was a 500 msec infrapitch signal. The infrapitch signal is made up of a repeated pink noise burst, for this test an infrapitch signal with a rate of 5 Hz was used.

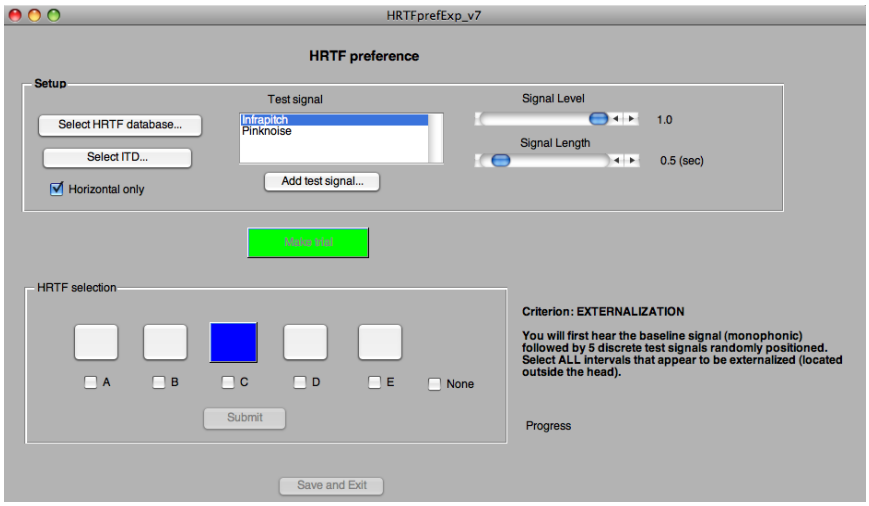

Figure 21.HRTF preference Graphical User Interface used during testing.

During each trial, test signals processed by 5 randomly selected HRTFs were presented, one for each interval (labeled A through E in Figure 21). As the interval was played to the subject, the button was highlighted. Subjects heard each interval at least once, but had the option of listening to the intervals as many times as necessary. They responded by selecting the checkboxes that corresponded to the intervals with stimuli that met the criterion of focus.

The test was performed in the Spatial Auditory Research Lab at New York University. The subject was seated at a table in front of a 17" MacBook Pro laptop with an Apogee "duet" audio

interface. Sennheiser HD650 open headphones were used to present the stimuli. All sources presented were static, where no head tracking was used during the experiment.

HRTF Datasets HRTF datasets were collected from two publically available databases: the database obtained at IRCAM and AKG from the Listen project and the CIPIC database measured at UCDavis. In addition, a dataset measured at the Naval Submarine Medical Research Laboratory (NSMRL) on KEMAR was included in the database. A total of 97 HRTF datasets were selected as the contenders for the HRTF selection process. All datasets were reformatted to conform to the NSMRL HRTF format.

The full database was scanned and reduced to a smaller, and more manageable, database of 27 HRTF datasets (13 from the IRCAM database, 13 from the CIPIC database, and the KEMAR dataset). This reduction was considered necessary in order to keep the experimental time reasonable and avoid subject fatigue. An objective measure was used to create the reduced database. The spectral contrast between front and back locations was used for frequencies 1000Hz to 10kHz. HRTFs with the highest spectral contrast were selected and used for the listening test.

Interaural time cues have a significant impact on the location perceived by the listener, especially at low frequencies. In order to eliminate any bias for HRTFs that may have ITDs similar to that of the subject, the ITDs contained in the HRIRs in the database were replaced by individualized ITDs measured on the subject taking the test.

Ten paid volunteers participated in the experiment. All subjects were students in the Music Technology program at New York University. All subjects had normal hearing. The experiment, including the HRTF measurement and listening test, took approximately 75-90 minutes. Each participant completed the experiment in one session.

Results

Experiments were performed to investigate whether individually measured HRTFs could be replaced by a user selection process and result in a percept that is as good as individually measured HRTFs. Results show that

1) Individualized HRTFs are chosen by most subjects as one of their preferred HRTFs, however,
2) There are several HRTF datasets that are selected almost as often as individually measured ones;
3) Many HRTF datasets were eliminated during the front/back discrimination phase, most likely due to the fact that the test was performed in a non-interactive setting;
4) There is agreement among subjects regarding which HRTFs are “better” or “worse”
5) There are different listener types – where a group of listeners prefer distinct, yet common, HRTF datasets from other listener types.

Results of this experiment indicate there is agreement among many subjects that there exist HRTF datasets that are selected almost equally often as the subjects’ individualized HRTFs. By

using a rudimentary selection procedure, we may be able to provide listeners with HRTFs that would provide very good cues, which would lead to an improved spatial auditory image in virtual environments while eliminating the high cost of measuring individualized HRTFs.

Analysis of the results for 10 subjects was performed, specifically focusing at how subjects selected their preferred HRTFs at each stage of the experiment. In the results presented below, the HRTF IDs represent the following HRTF sets:

- #1: individualized HRTF
- #2 - #14: IRCAM datasets
- #15 - #27: CIPIC datasets
- #28: KEMAR HRTF

Externalization Results for the externalization stage of the listening test are presented in Figure 22. The figure shows the percentage of subjects that chose each of the 28 HRTFs presented for the "externalization" criterion. Only HRTFs that were chosen with at least a 67% confidence level (2 out of the 3 times each HRTF was played) were considered reliable and are shown in the figure. Results show that 90% of subjects selected their individualized HRTF (ID#1) as being perceived to have an externalized sound image. The CIPIC datasets and the KEMAR dataset were chosen by 70-80% of subjects for their externalization quality. In contrast, the IRCAM datasets were selected by 10-40% of subjects.

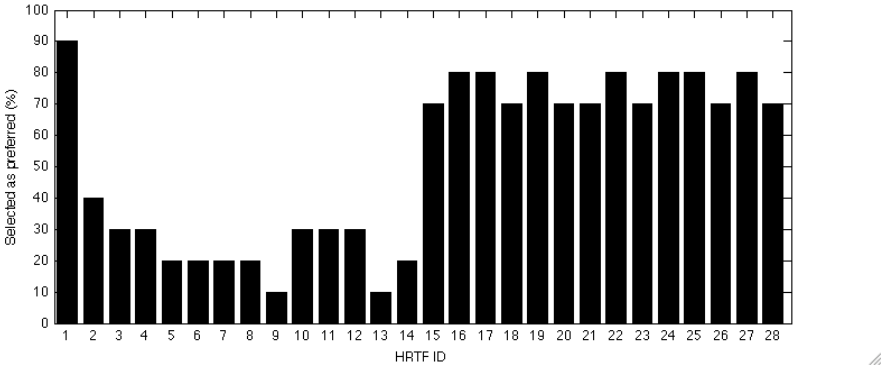

Figure 22. Bar graph of the percentage of subjects selecting each HRTF in the database during the first "externalization discrimination" stage of the experiment

Elevation discrimination The listening test's second stage results, where subjects were asked to discriminate high from low elevations, are shown in Figure 23. Similarly to the results from the first stage, 90% of subjects selected their individualized HRTF. However, there are several datasets from the CIPIC database where we see a selection rate of up to 80% (e.g. HRTF ID #24).

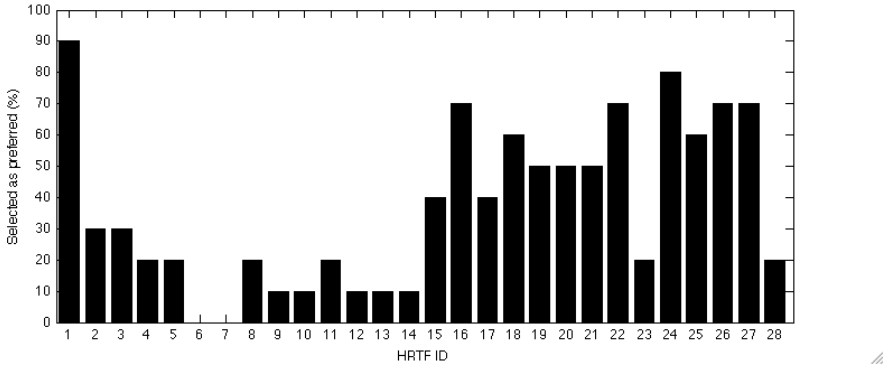

Figure 23. Bar graph of the percentage of subjects selecting each HRTF in the database during the second "elevation discrimination" stage of the experiment

Front/back discrimination Selection results from the third stage, front/back discrimination, are presented in Figure 24. Many HRTFs were eliminated at this last stage of the listening test. The selection of the individualized HRTF drops to 70%. We observe one HRTF dataset (ID#27) that has a selection rate of 60%, while others have been only selected 40% times, or less. Several of the datasets from the public databases have been eliminated.

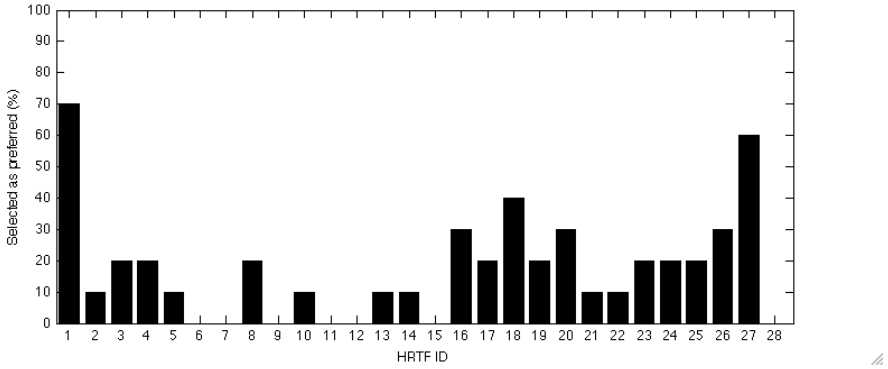

Figure 24. Bar graph of the percentage of subjects choosing each HRTF in the database during the third "front/back discrimination" stage of the listening test . The results shown are in effect the "winning" HRTFs.

Research shows that front/back confusion rates increase when non-individualized HRTFs are used to simulate spatial audio. This is particularly true in non-interactive environments. Although in this case subjects were not asked to judge the absolute location of a source, but rather to compare between two sources presented along the cone of confusion on the horizontal plane, the source of the significant number of rejections could be due to the fact that discrimination was performed in a static environment.

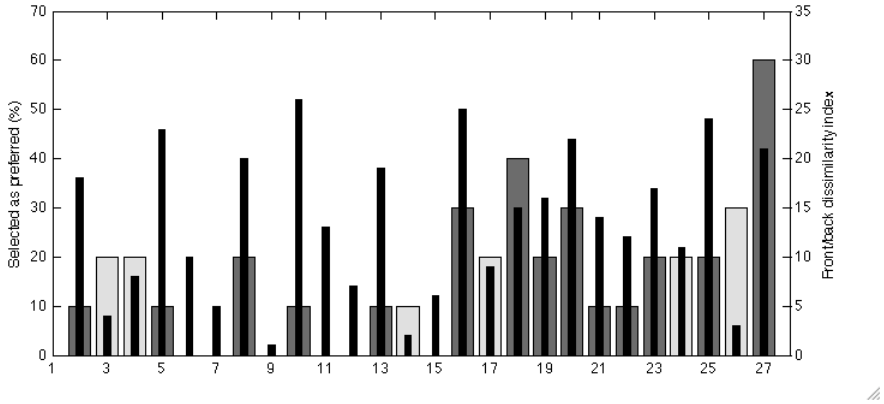

Figure 25. Comparative bar graph of the percentage of subjects choosing each HRTF during the front/back discrimination stage (light and dark grey) and the front/back spectral dissimilarity rank (black) used to pre-select the datasets. Top 15 ranking dissimilarity HRTFs (with highest dissimilarity index) are highlighted in dark grey.

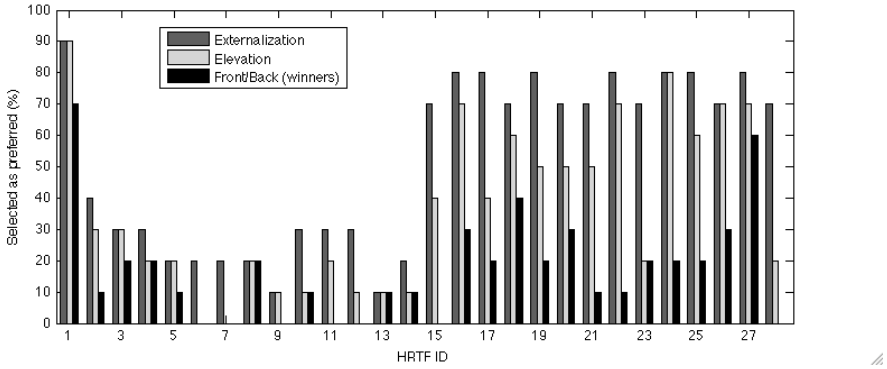

Figure 26. Results of all 3 stages of the experiment: externalization (dark grey), elevation discrimination (light grey) and front/back discrimination (black)

As mentioned above, the measure used to pre-select the HRTF datasets from the full publically available databases to the reduced 27 HRTF database was the amount of spectral dissimilarity between front and back locations along the cone of confusion on the horizontal plane. Figure 25 presents a comparative bar graph with the percentage of subjects selecting each HRTF and the dissimilarity ranking based on the dissimilarity index calculated above. As can be seen, there is a high correlation between the HRTFs containing a higher spectral dissimilarity content and those selected by the subjects in the listening test. HRTFs with a lower dissimilarity index (and lower dissimilarity ranking) were typically not selected in this final stage of the listening test.

Subject groups A compilation of the selection results of all 3 stages is presented in Figure 26. In this figure it appears that HRTFs with IDs 2-14 (IRCAM database) are generally selected much less often than are HRTFs with IDs #15-27 (CIPIC database). At first, it may appear that these are typically less popular and less preferred HRTFs. However, analysis of the selection of each subject shows that there is another reason.

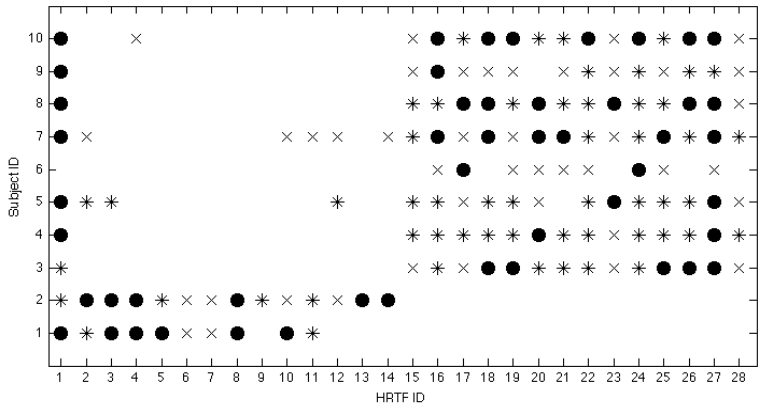

Figure 27. HRTF selection preference for each subject, for the three stages of the listening test: externalization (x), elevation discrimination (*), and front/back discrimination (filled circles).

Results of the selection process for each stage, for each subject, are shown in Figure 27. Externalization selections are presented with an 'x' marker, HRTFs selected after the externalization and elevation discrimination phase are shown with a star, final winners are presented using filled circles. We observe that selection results are much different for subjects 1 & 2, and subjects 3-10. With a few exceptions, subjects 1 & 2 generally tended to choose HRTFs from the IRCAM database, while subjects 3-10 had a strong preference for the CIPIC database and KEMAR measurements. This indicates that there may be different types of listeners, with preferences for distinct groups of HRTFs datasets.

Phase II

Based on the results of the first experiment described above, a preselected bank of 10 most preferred HRTFs was selected for further study. In addition, each subject's personalized HRTFs were included in the listening test.

Similarly to the first phase, this experiment consisted of two parts: HRTF measurement of the subject, and a listening test.

HRTFs were measured at the Spatial Auditory Research Lab at New York University using the same setup and equipment as described above.

The task during the listening test asked the subjects to select all the HRTFs that fulfilled the criterion they were listening for. The listening test was divided into three stages. At each stage a different criterion was the focus of the listening task. Subjects were asked specifically to pay attention to the criterion and select all the intervals where the criterion was fulfilled. The following criteria were used:

- Stage 1 – Externalization: source appears to be located outside the head.
- Stage 2 – Elevation discrimination: capability to discriminate a source located at a high and low elevation, given a fixed azimuth.
- Stage 3 – Front/back discrimination: capability to discriminate a source located in front from one located in the back, along the cone of confusion.

The search for user-selected HRTFs giving a good spatial image was not driven by making judgments about the absolute source location. The goal was to make judgments about the spatial impression of an image relative to a reference signal. Each stage of the test presented the subject with a different reference signal. The subject was asked to make a judgment based on the criterion under study.

HRTFs Eleven HRTF datasets were used during the test: the individually measured HRTF and 10 datasets selected from publically available databanks. Public HRTFs were taken from the database obtained at IRCAM and AKG from the Listen project, and from the CIPIC database measured at UCDavis.

The public datasets used during this experiment were selected based on the results obtained during the study described above. From the 27 HRTF datasets (not including individualized HRTFs), the 10 most preferred HRTF sets were selected and used in this study. Table 2 contains the list of the HRTFs used during the experiment.

Table 2. HRTFs used in experiment

HRTF	HRTF	Subject
1	Individually	
2	LISTEN	1014
3	LISTEN	1022
4	LISTEN	1028
5	CIPIC	12
6	CIPIC	15
7	CIPIC	27
8	CIPIC	58
9	CIPIC	119
10	CIPIC	131
11	CIPIC	154

Interaural time difference (ITD) cues are one of the principal cues in determining the location of sound sources on the horizontal plane, particularly at low frequencies. These cues alone have a very strong impact on perceiving a source's location. The public HRTF datasets contained ITDs measured on the individual subjects. In order to eliminate fundamental differences in location perception, as well as any bias for HRTFs that may have ITDs similar to those of the subject, all ITDs in the public datasets were replaced by individually measured ITDs.
Test signals Three test signals were used to study the stimulus dependence of HRTF user selection: Infrapitch, Speech, and Music signals. The stimulus used in this study used a sample of pink noise which was repeated at a rate of 5 Hz for 500 msec.

The Speech signals were recordings of a male voice speaking numbers from zero to nine. Each number was recorded individually, and was less that 350 msec in duration. During the listening

test, numbers were selected randomly for each trial. For each interval within a trial, the numbers were the same.

The Music test signals were recordings of arpeggiated triads played on a piano. Chords from C6 (fundamental frequency 1046.5Hz) to C7 (fundamental frequency 2093Hz) were recorded. Similarly to speech signals, the music signals for each trial were selected randomly. Within a trial, the music signals were kept consistent for all intervals.

Twenty paid volunteers participated in the experiment. Subjects were recruited in the Music Technology program at New York University. All subjects had normal hearing. The two parts of the experiment took 75-90 minutes. Each participant completed the experiment in one session.

Criteria-specific results

Externalization Averaged results for the externalization stage, for all three signals, are presented in Figure 28. The figure presents the percentage of subjects who chose each HRTF as giving a good percept of externalization, when compared to a monophonic reference signal. None of the HRTF sets were chosen unanimously. Results show that two HRTFs were chosen equally often by the same number of subjects as most preferred: the subject's own individually measured set, and HRTF set #5. Both were selected by 90% of the subjects. In other words, 18 out of the 20 subjects who took the test selected their individually measured and HRTF #5 as giving a good perception of externalization. Several other HRTFs were selected by over 70% of the subjects.

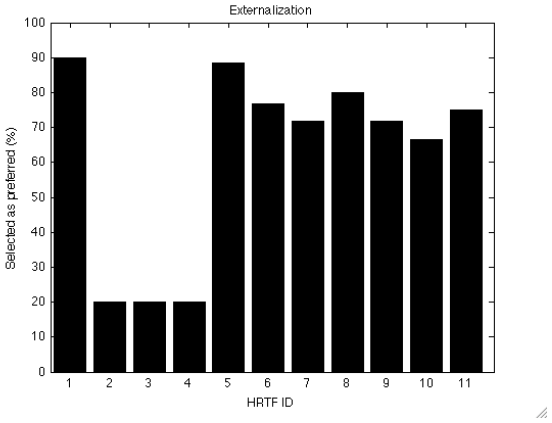

Figure 28. Percentage of subjects selecting each HRTF in the database during the externalization discrimination stage of the experiment.

Results for the externalization stage for each of the three test signals are presented in Figure 29.

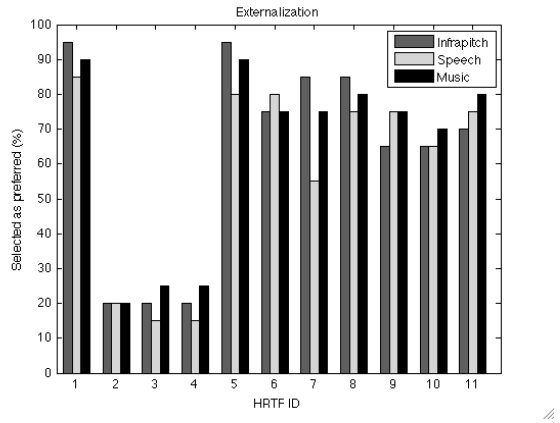

Figure 29. Comparative graph of selectivity preference for externalization, for each stimulus.

Elevation discrimination Figure 30 presents average preference results from the elevation discrimination stage of the listening test. The highest ranked HRTF is #5, which was selected by 65% of subjects. The individualized HRTF, and HRTFs #8 and #9 were the next most preferred, being selected by 60% of subjects. HRTF #11 was selected by 53% of subjects. All others had a selection rate of 40% or less.

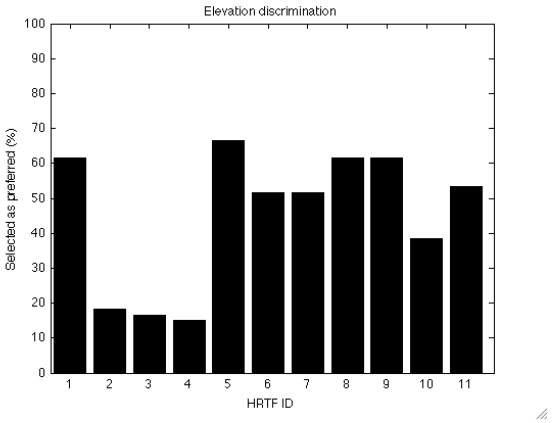

Figure 30. Percentage of subjects selecting each HRTF in the database during the elevation discrimination stage of the experiment

Figure 31 breaks down the preference by test signal. As can be seen in this figure, the Infrapitch signal (which contains an overall broader spectrum than the other two types of test signal) resulted in a greater number of subjects selecting many of the presented filter sets.

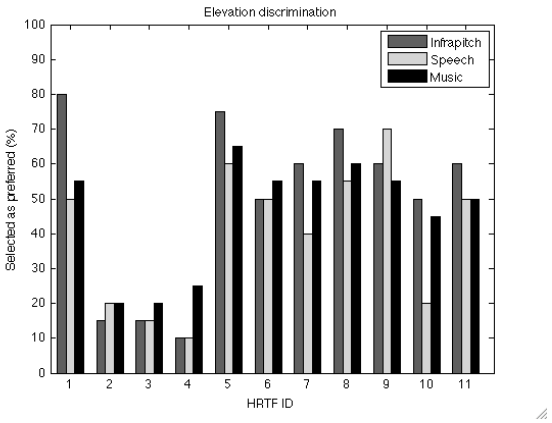

Figure 31. Comparative graph of selectivity preference for elevation discrimination, for each stimulus.

Front/back discrimination During the final stage of the test the selection of the individualized HRTF drops to 50%. Except for HRTF #5 (with a selection rate of 55%), all other datasets were selected by fewer than 50% of subjects, with selection rates between 10%-40%. Average front/back discrimination results for all test signals are presented in Figure 32, and stimulus-dependent results are shown in Figure 33.

These results are consistent with the Phase I experiment, and are expected.

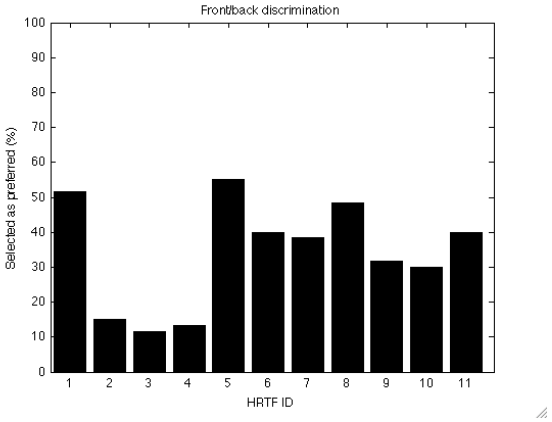

Figure 32. Percentage of subjects selecting each HRTF in the database during the front/back discrimination stage of the experiment

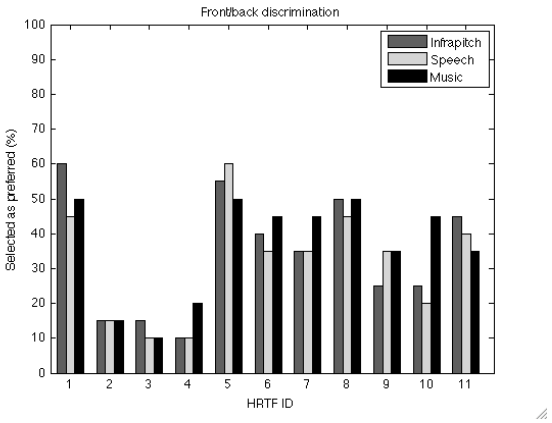

Figure 33. Comparative graph of selectivity preference for front/back discrimination, for each stimulus.

Table 3. Preference ratings for the Infrapitch signal

Externalization			**Elevation discrimination**			**Front/back discrimination**		
RANK	*HRTF ID*	*% SEL*	*RANK*	*HRTF ID*	*% SEL*	*RANK*	*HRTF ID*	*% SEL*
1	1 (Individualized)	95%	1	1 (Individualized)	80%	1	1 (Individualized)	60%
1	5	95%	2	5	75%	2	5	55%
2	7	85%	3	8	70%	3	8	50%
2	8	85%	4	7	60%	4	11	45%
3	6	75%	4	9	60%	5	6	40%
4	11	70%	4	11	60%	6	7	35%
5	9	65%	5	6	50%	7	9	25%
5	10	65%	5	10	50%	7	10	25%
6	2	20%	6	2	15%	8	2	15%
6	3	20%	6	3	15%	8	3	15%
6	4	20%	7	4	10%	9	4	10%

Table 4. Preference ratings for the Speech test signal

Externalization			Elevation discrimination			Front/back discrimination		
RANK	*HRTF ID*	*% SEL*	*RANK*	*HRTF ID*	*% SEL*	*RANK*	*HRTF ID*	*% SEL*
1	1 (Individualized)	85%	1	9	70%	1	5	60%
2	5	80%	2	5	60%	2	1 (Individualized)	45%
2	6	80%	3	8	55%	2	8	45%
3	8	75%	4	1 (Individualized)	50%	3	11	40%
3	9	75%	4	6	50%	4	6	35%
3	11	75%	4	11	50%	4	7	35%
4	10	65%	5	7	40%	4	9	35%
5	7	55%	6	2	20%	5	10	20%
6	2	20%	6	10	20%	6	2	15%
7	3	15%	7	3	15%	7	3	10%
7	4	15%	8	4	10%	7	4	10%

Table 5. Preference ratings for the Music test signal

Externalization			Elevation discrimination			Front/back discrimination		
RANK	*HRTF ID*	*% SEL*	*RANK*	*HRTF ID*	*% SEL*	*RANK*	*HRTF ID*	*% SEL*
1	1 (Individualized)	90%	1	5	65%	1	1 (Individualized)	50%
1	5	90%	2	8	60%	1	5	50%
2	8	80%	3	1 (Individualized)	55%	1	8	50%
2	11	80%	3	6	55%	2	6	45%
3	6	75%	3	7	55%	2	7	45%
3	7	75%	3	9	55%	2	10	45%
3	9	75%	4	11	50%	3	9	35%
4	10	70%	5	10	45%	3	11	35%
5	3	25%	6	4	25%	4	4	20%
5	4	25%	7	2	20%	5	2	15%
6	2	20%	7	3	20%	6	3	10%

Signal-dependent results Results showing selections for the 3 different test signals at each stage of the discrimination are presented in Table 4. Preference ratings for the Speech test signal, Table 4 and Table 5. The tables show the percentage of subjects that selected each HRTF at every stage of the listening test, as well as their rank. Figure 34, Figure 35, and Figure 36 contain the corresponding graphical representation of these results.

These results show that none of the HRTFs, including individually measured, were selected unanimously. However, several datasets were selected by the majority of subjects.

The Infrapitch test signal (which contained the broadest spectrum of all test signals) produced higher selection rates than other test signals, with some datasets being selected by as many as 95% of subjects for externalization, 80% for elevation discrimination and 60% for front/back discrimination. The Speech test signal produced overall lower selection rates across all HRTFs, with a highest value of 85% for externalization, 70% for elevation and 60% for front/back discrimination. The Music test signal resulted in a selection rate of 90% for externalization, 65% for elevation and 50% for front/back discrimination.

The results are interesting from several perspectives. First, the individualized HRTF is not the most agreed upon preferred set for every test signal, at every stage. Second, there appear to be several datasets that are selected by many subjects for one criterion, but lose much of the selectivity rating for another criterion. For example, the individualized HRTF for the Speech signal was selected by 85% of subjects for the externalization quality, but only by 50% of subjects for elevation discrimination – a 35% decrease. Conversely, there are datasets that did not lose as much from one stage to another. For example, in the Speech signal, HRTF #9 was selected by 75% of subjects for externalization, and by 70% for elevation discrimination – a 5% decrease only. From this, we can say that there may be HRTFs that are preferred by more subjects when listening to a specific criterion, but that may not work as well for all those subjects when a different criterion is considered.

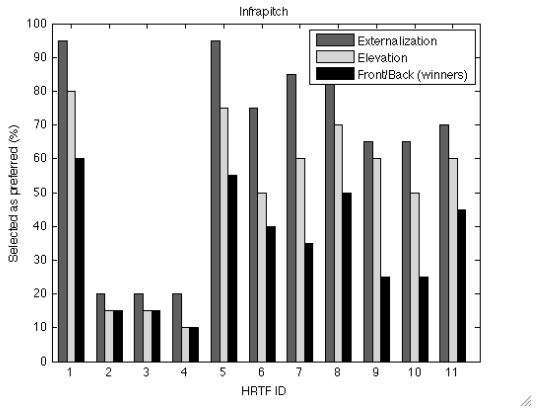

Figure 34. Preference results at each stage of the listening test for the Infrapitch test signal

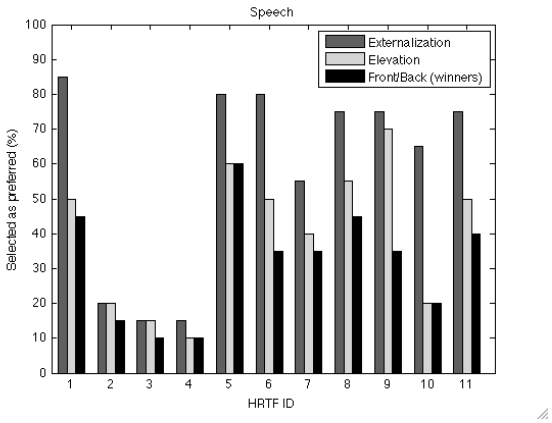

Figure 35. Preference results at each stage of the listening test for the Speech test signal

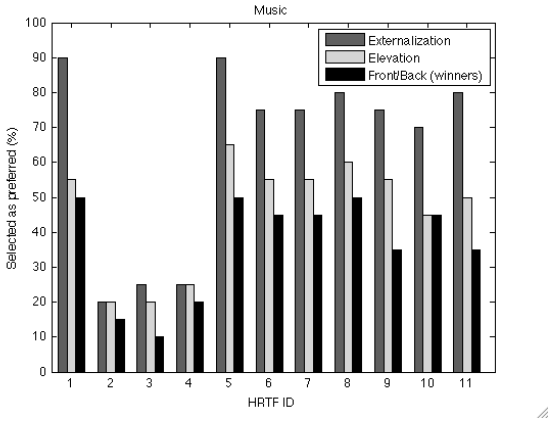

Figure 36. Preference results at each stage of the listening test for the Music test signal

Listener groups Results in the second phase of this line of research concur with the results from Phase I, where we saw evidence of listener groups that preferred a subset of HRTF datasets. Here, three of the HRTF datasets (#2, #3, #4) were selected by fewer subjects than other datasets. For all test signals, at every stage of the listening test, these three datasets were selected by no more than 25% of subjects. Figure 37, Figure 38 and Figure 39 present the selected HRTFs (with a 67% confidence rating) by each subject for every HRTF dataset, for the Infrapitch, Speech, and Music test signals, respectively. In these figures, an HRTF that was selected by a subject during the externalization stage only is shown by an 'x' symbol; an HRTF selected during the externalization and elevation discrimination stages is represented by an asterisk; an HRTF selected during all three stages is shown with a solid circle.

These results show that a small subset of listeners selected the HRTF datasets that were generally less liked by the overall subject population. For example, in Figure 37, Subject #11 selected HRTFs #2, #3 and #4, in addition to their individualized HRTF, #6 and #8, to be among their favorite ones. A similar pattern can be seen for subjects #8, #13, and #14.

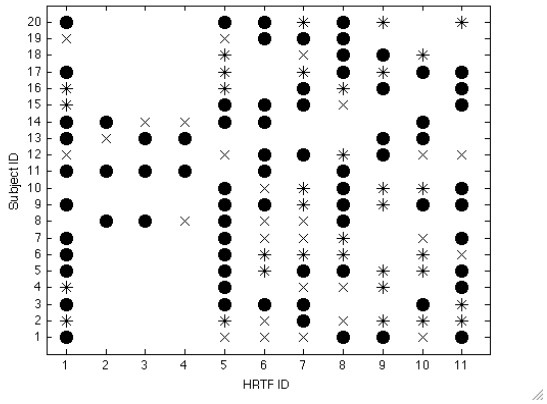

Figure 37. Subject-dependent HRTF selection preference, for all stages, with the Infrapitch signal: externalization (x), elevation discrimination (*), and front/back discrimination (filled circles).

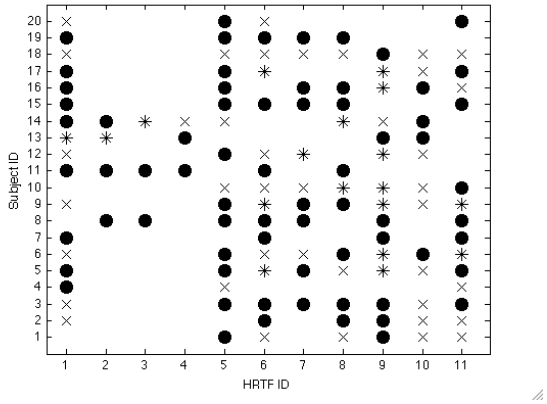

Figure 38. Subject-dependent HRTF selection preference, for all stages, with the Speech signal: externalization (x), elevation discrimination (*), and front/back discrimination (filled circles).

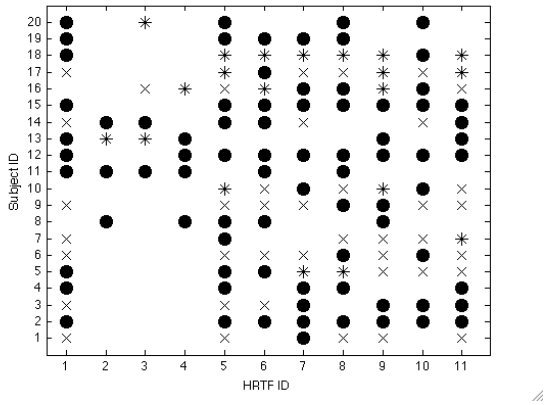

Figure 39. Subject-dependent HRTF selection preference, for all stages, with the Music signal: externalization (x), elevation discrimination (*), and front/back discrimination (filled circles).

Discussion

Individualized HRTFs were selected by most subjects in at least two stages of the experiment. These results are compliant with previous research indicating better localization performance with individualized HRTFs.

During the front/back discrimination stage of the experiment, many HRTF datasets were eliminated, indicating that front/back discrimination may be the more complex dimension to successfully synthesize using non-individualized HRTFs. However, it is important to note that the test was conducted in a static environment where the subject had no interaction via head-tracker or otherwise. Previous research (Begault et al, 2001) has shown that the incorporation of head tracking reduces front/back confusions. Although this experiment did not test absolute localization (only relative judgment was tested), we predict that the lack of interaction with the environment may still be the grounds for these results.

There are selected HRTFs that come very close to the selection preference of the individualized HRTFs. In particular during the externalization and elevation discrimination stages, several HRTFs from the public databases were selected almost as often as the individualized one. This leads us to believe that there exist HRTF datasets that are preferred by a significant number of listeners.

Results have shown that there are groups of subjects that prefer distinct sub-groups of HRTFs. In this experiment, subjects were clearly divided between the databases of HRTFs they preferred. This proves that there are different "categories" of listeners. Although there is evidence that there are subjects who preferred a sub-group in one of the databases, these findings are not significant enough due to the limited number of subjects. Although a "one size fits all" HRTF may not exists due to the great disparity in body and ear shapes and sizes, results of this experiment indicate there is agreement among many subjects that, aside from front/back discrimination, there exist HRTF datasets that are selected almost equally often as the subjects' individualized HRTFs. By using a rudimentary selection procedure, we may be able to provide listeners with HRTFs that, although not as ideal as individualized HRTFs, would provide very good, and certainly better than generic, cues which would lead to an improved spatial auditory image in virtual environments while eliminating the high cost of measuring individualized HRTFs. The results of the second phase of the study confirmed that not all HRTFs are good for every listener. Results showed that there are some datasets that are selected by a large number of subjects for a specific stimulus and criterion, but when subjects focused on a different criterion the selectivity rating dropped significantly. This indicates that some datasets may be better suited for providing a better perception of specific aspects of spatial audio. Conversely, there are datasets that offer a compromise, and perform "pretty well" for all criteria.

Results of the first phase of the study were presented in a publication at the UHSI Symposium in Providence, RI July 27-29 (Roginska et al., 2010c). Results from the

second phase will be presented in a publication at the 129th Audio Engineering Society Convention in San Francisco, CA, November 4-7, 2010 (Roginska et al., 2010a).

Topics for further investigation

Our research has shown that it is possible to preserve several spatial attributes of sound when rendering auditory environments over HRTFs different from those of the listener. Such rendering still requires some form of individualization: user-specific ITDs were employed in the experiments and not all listeners preferred the same sets of HRTFs. The results also depend on the choice of auditory scene. We strongly recommend a research and development program that determines the importance of the many user-dependent factors that still remain. This research program should also engage more relevant operational tasks to determine the quantitative effects on performance of substituting even more generic options for those included in the present study.

IMPROVED PERCEPTUAL SIGNAL-TO-NOISE RATIO IN SPATIAL AUDIO METHODS

Ambient sea-state noise presents a constant acoustic background in underwater environments against which targets of interest are perceived. When simulated in a spatial virtual auditory environment, targets at specific locations may become partially or fully masked due to the additive effects of such background noise coming in from all directions. Standard approaches to enhancing the target either are ineffective or introduce distortion in either the target or the background noise, thereby potentially altering the perceived nature of these sources.

Listening Everywhere through Two Ears

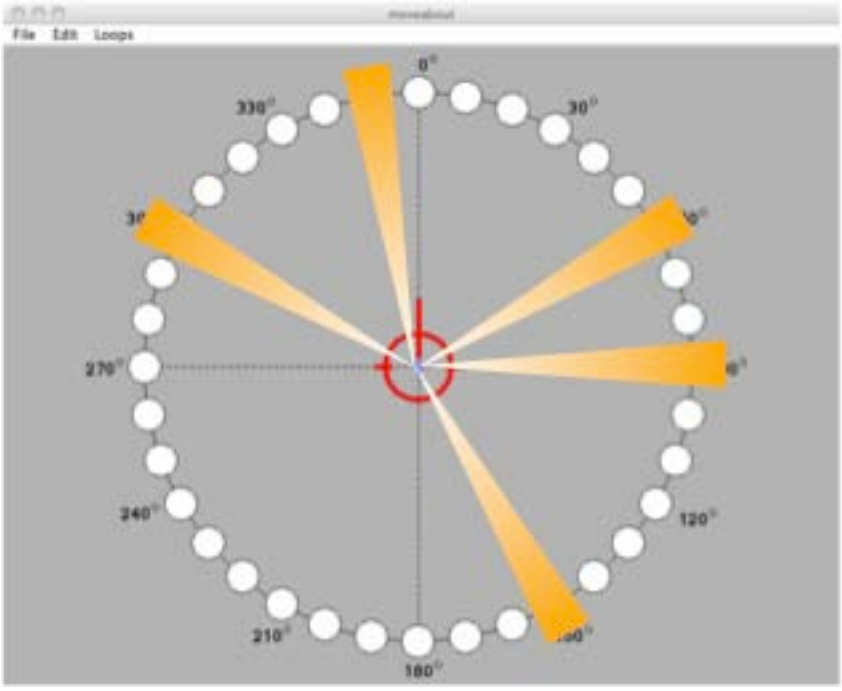

Figure 40. A schematic top-down perspective illustrating the ability to monitor a number of different locations simultaneously. Performance appears constrained by the inherent limits of steering a two-sensor array, rather than cognitive-decision factors.

The Detection experiments conducted during Year 1, along with the Dual Task experiments conducted during Year 2, support the claim that operators are capable of listening, in parallel, along a number of different bearings. Nevertheless, operators are still constrained by the mathematical limits of two sensors, e.g., their ears, which impose bounds on the degree to which acoustic events present in other beams can be suppressed while listening along a particular beam. That is, at a conceptual level, the operator establishes a family of parallel processes to steer a two-sensor array in a number of different bearings simultaneously. By monitoring the outputs of these parallel processes, the operator gains better situational awareness than is possible with sequential listening in single directions, but their ability to detect acoustic events along any particular bearing is limited by the signal-to-noise ratio of the steered array, rather than that along a single bearing. Accordingly, the operator can listen everywhere all the time, but what he or she hears in any given direction is limited by the noise and clutter coming from all directions.

Methods

Coupling Binaural Perception and Auditory Stream Formation Our current understanding of auditory perception is that the creation and maintenance of auditory streams plays a central role in maintaining our representation of the acoustic world around us. Often, but not always, there is a direct correspondence between streams and acoustic sources. When a direct correspondence breaks down, the auditory stream tends to “fill in” or “complete” the properties of the source even when the source has been eliminated or altered. One

way to think about this process is to imagine a pre-perceptual sensory state filled with atoms of sound – short snippets of sound that may be localized in time and frequency – and that stream formation is a process of associating the atoms within and across frames of time. Once formed, these organized atoms emerge as perceptually well-formed auditory streams. If the acoustic source providing the atoms is substantially altered, or disappears altogether, it is still possible, under this formulation, for the auditory stream to persist unaltered as long as the proper collection of sound atoms are provided by some other acoustic source.

In the case of binaural perception, particularly with multiple sources of noise, this simple model of auditory stream formation suggests that we can eliminate atoms from noise at a given azimuth without impairing our perception of that noise as long as the stream-formation process finds other atoms that are consistent with the noise stream. Because we are limited to hearing through two ears, our system is inevitably unable to delineate the origins of many atoms such that source continuity is maintained by an ever-abundant supply of atoms from surrounding noises in the acoustic field. The underwater environment would appear to be a perfect generator of such noise atoms.

Granularization is a signal-synthesis technique that originated in the computer music community. It owes its intellectual heritage to the work of Gabor (1946) and his interest in precisely defining the "time-frequency atoms" out of which sound or other signals can be constructed. It is also intimately related to a host of signal processing techniques with more formal labels, such as wavelet decomposition and frames, all of which are applicable to problems of noise reduction. Our problem, however, is not quite that of noise reduction in the statistical signal processing sense, but of figure-ground enhancement in more of the music composition/perception sense. That is, the problem in spatial audio for underwater applications is to improve *both* the definition of the figure (a target) and the ground (the background noise). Eliminating the noise altogether, or substantially altering the perceptual qualities of either the figure or the ground, introduces artifacts into our spatial perception, which are unacceptable for many underwater applications. We, therefore, refer to our work as one of granularization, to emphasize the notions of perceptual figure and ground.

Structurally, granularization synthesizes sounds from atoms. In a figure-ground case, we are given a set of atoms and must decide which atoms belong to the figure and which to the ground. Because "good" figures and "good" grounds exist as auditory streams, we must also decide which atoms are sufficiently ambiguous that they don't lend themselves to supporting "good" versions of either. Similarly, we must decide which atoms are interfering (or masking) those atoms that are integral in constructing a particular stream. Once these decisions are made, granularization provides the means for reconstituting the audio environment by synthesizing it from the atoms we've selected.

In our studies, we have been concerned primarily with the broadest version of this granularization picture. There are numerous details, which are likely to improve or degrade the performance of the technique. Nevertheless, if the principles of auditory streaming can be employed as described above, our hypothesis is that even the most

straightforward method for decomposing the complex sound into its atoms should show some promise after proper reconstitution.

We employed a nearly perfect reconstruction filter-bank to decompose the input noise at any location into narrowband “molecules”. To pick out the atoms from these molecules, we employed a technique that has found recent application in the visual perceptual-lossless compression domain. This technique uses a deadband threshold to map expansion coefficients in a data compressor to zero. When applied to sound, this is equivalent to setting an amplitude threshold at the output of the narrowband filter and mapping all values below that threshold to zero. The narrowband signals that remain are the atoms we wish to preserve. These steps are outlined in Figure 41 below and described more completely in the caption.

As we explored this implementation of the granularization idea, we discovered that it was possible to eliminate a relatively large proportion of noise atoms at any one location (84%) without degrading the perception of the noise field itself. This result is consistent with the auditory streaming hypothesis: the significant loss of noise atoms at any one location is offset by the presence of atoms that lie above the deadband threshold at other locations. Because the auditory system fuses all of the different noise sources into the percept of a field of background noise, the “holes” are filled and the percept of a continuous, dense noise persists.

We also discovered that as we eliminate more of the noise atoms in any one beam, the proportion of atoms remaining are more likely to contain a target, if one should be present in that beam. While preserving the “good” ground of the noise field, we believe that this simple algorithm removes those atoms that mask or blur the “good” figure of the target. Pilot data demonstrated some improvement in detection threshold for the target. Based on those data, we conducted a more formal experiment.

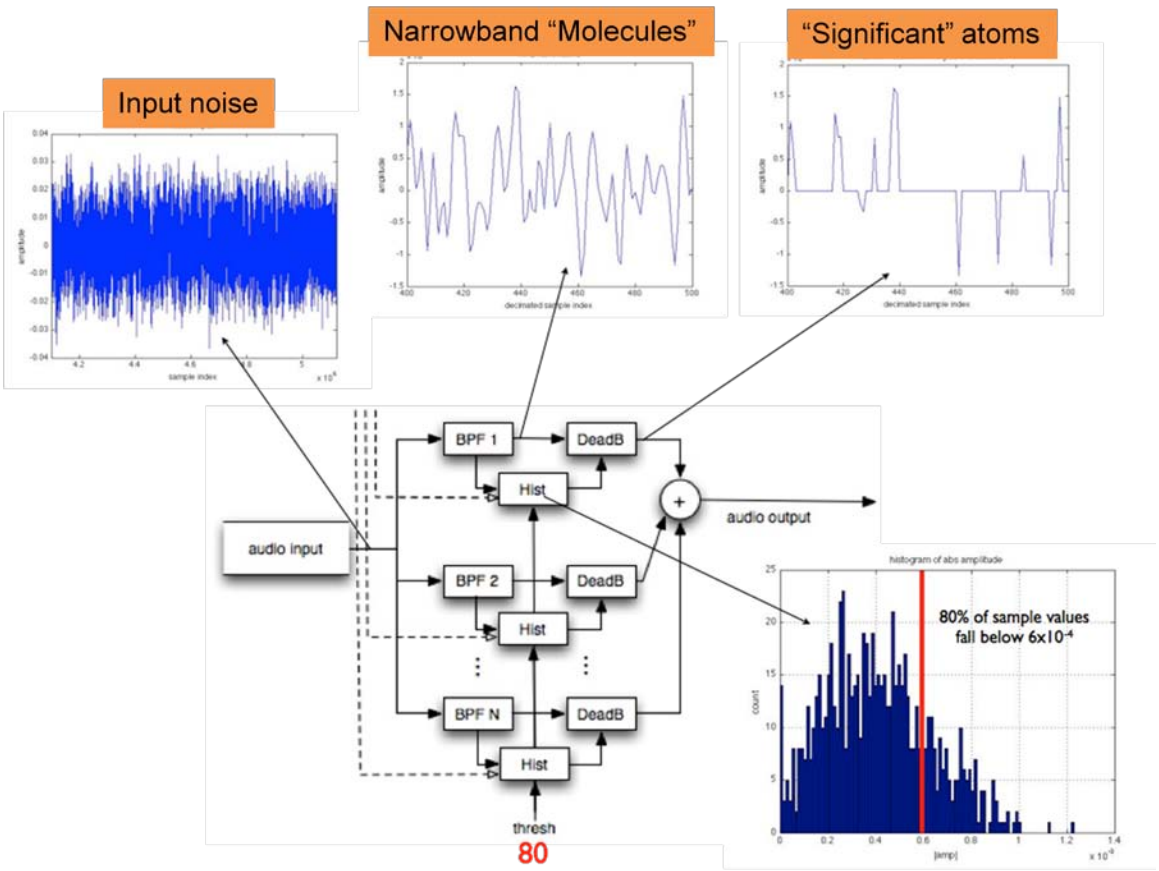

Figure 41. A block diagram of a simple implementation of granularization is shown, along with several signal taps. Audio input is passed through a filterbank. Histograms of the narrowband output amplitudes are estimated from which the value of the deadband threshold is calculated. Significant "atoms" are culled from each collection of narrowband "molecules" by mapping the amplitude of each narrowband signal to zero when its value fell below the value of the deadband threshold. Typical thresholds lie in the range of 50 to 99%, which means that up to 99% of the narrowband signal amplitude is mapped to zero by the thresholding operation.

Experiment A psychophysical 2IFC Levitt procedure was used to measure the detection threshold of a sinusoid in noise. The noise was either at the same location as the target (0 deg. azimuth at the listener's midline), called the onbeam masker condition, or 35 noises were present every 10 degrees in azimuth in addition to the onbeam noise. The latter was called the circle mask condition. The duration of the sinusoid was 1 second. Thresholds for three different frequencies were measured: 1, 2, and 4 kHz. The experiment was repeated after all beams had been granularized with a deadband threshold of 84%.

Seven music technology students at New York University served as subjects. All had clinically normal hearing and were highly-experienced in listening to audio. Individualized HRTFs were used to spatialize the data using a measurement system as described above in the section on Auditory Search.

Results

A simple version of an approach that takes advantage of stream formation processes in audition to mitigate threshold loss in spatial audio was examined. Results from five subjects show that:

1) It is possible to recover 6 of the 10 dB elevation in detection threshold for sinusoidal targets in an environment of 36 spatially distributed pink-noise sources.
2) This recovery appears to be frequency-dependent, with nearly perfect recovery occurring at the highest frequency tested (4 kHz).
3) This improvement is achieved without degrading the perceptual quality of the pink-noise sources.

The results establish the proof-of-concept that properties of both the underwater acoustic environment and the human auditory system can be capitalized upon to enhance the perceived figure-ground in a spatial audio display, which supports situational awareness.

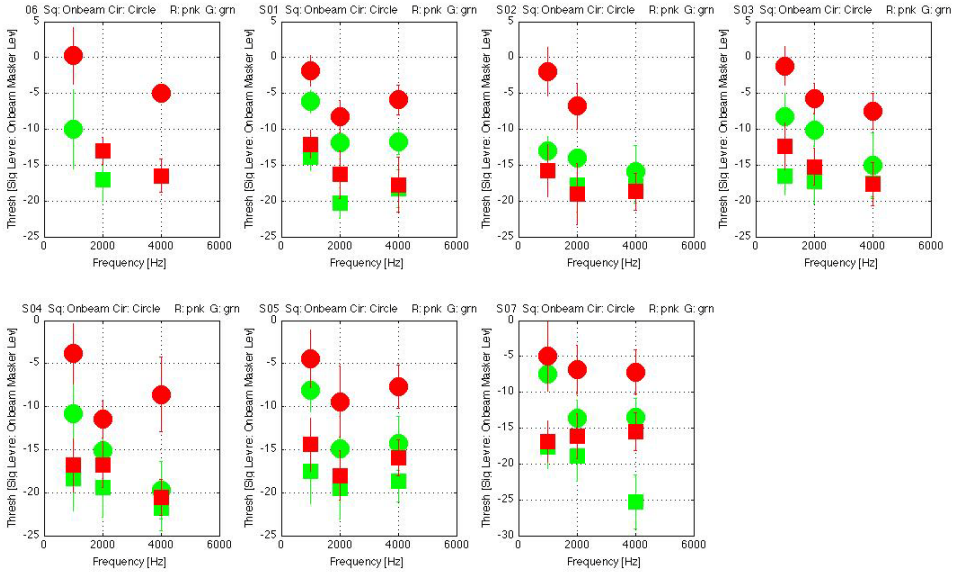

Figure 42. Data from seven subjects are shown in the panels. The acoustic treatment of the sounds is indicated by color (unprocessed: Red, granularized: Green). The masking condition is indicated by symbol (onbeam: Square, circle: Circle). The frequency of the sinusoid is indicated in Hz along the abscissa and the SNR of the detection threshold is indicated in dB along the ordinate. Thresholds are based on three blocks of the Levitt tracking procedure. Error bars show ±1 standard deviation about the mean.

Figure 42 above shows the results from the seven subjects. Each panel shows the detection thresholds (ordinate) for the different experimental conditions as a function of the frequency of the sinusoid (abscissa). Each threshold is based on three measures of detection threshold, using the Levitt procedure with a two-down/one-up rule. Data for

two of the seven subjects were incomplete and they were removed in computing average performance (shown in Figure 43 below).

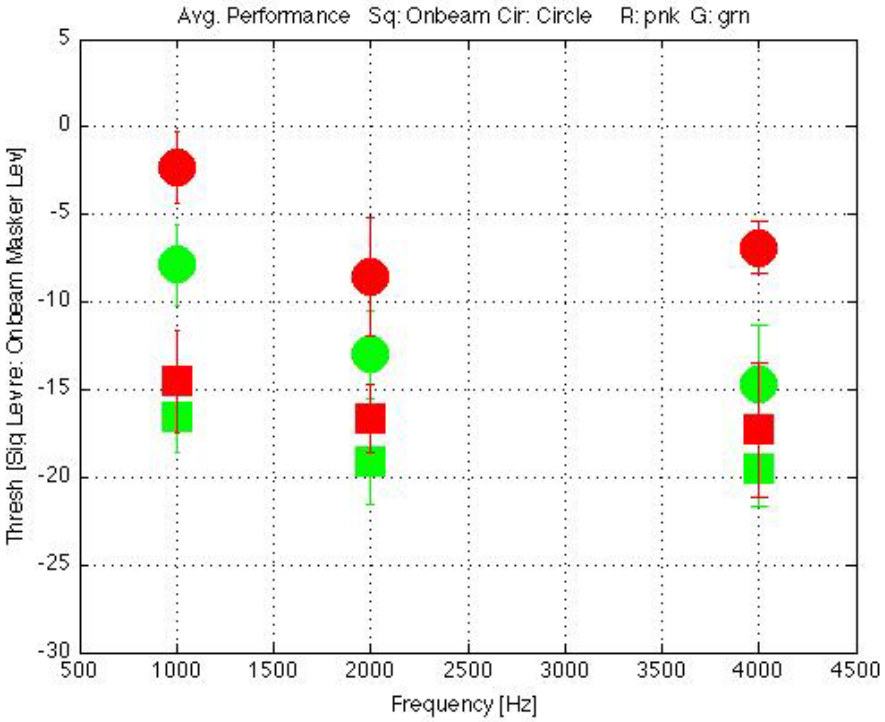

Figure 43. Average data across five subjects who had complete sets of data. Same symbol and color code as found in Figure 42.

When averaged over the five remaining subjects, we see clear evidence of an effect of granularization on threshold. The red symbols denote the detection thresholds for the unaltered acoustic condition whereas the green symbols denote thresholds for the granular-processed acoustic condition. The squares and circles correspond to the onbeam masking and circle masking conditions, respectively. Comparing the red circles to red squares, we see that the addition of 35 maskers in this spatial audio environment elevates detection thresholds by an average of 10 dB. Granularization of the circle masking condition recovers, on average, 6 dB of that threshold loss. There also appears to be an interaction between frequency and granularization: for the 4 kHz case, the amount of unmasking from granularization almost eliminates the masking effect of the additional 35 noise sources.

Discussion

The results also show that granularization in the onbeam condition improves detection threshold. On average, this improvement is 2.25 dB. Nevertheless, this gain comes at the price of completing destroying the enhanced figure-ground of the tone and noise: at a level of 84%, listeners clearly hear an artifact in the noise which some described as *chatter*. Without having additional noise atoms to draw from in the circle masker condition, the removal of so many noise atoms dramatically changes the perceptual nature of the ground against which the target is heard.

In further pilot testing, we have determined that further reduction in threshold is possible in the case of circle maskers, but these come at a similar price of introducing granular chatter to the noise field stream. We have also observed 8 dB or greater improvements in threshold when using wideband noise as the target. This is consistent with the frequency effect observed with respect to sinusoids and supports the idea that the auditory system is constructively creating an auditory stream for a noise target from the atoms that are available within the binaural pre-perceptual representation.

Besides the very simple algorithm used to decompose the acoustic signals into atoms and decide which ones to remove, the experiment also over-simplifies the acoustic environment and listening task by modeling targets as pure tones and background noise as stationary pink. Nevertheless, we believe our findings are a proof of the concept that one can take advantage of the additional noise that comes when listening to a fully spatialized underwater environment to recover some of the detection loss.
Results of this study were presented in a publication at the UHSI Symposium in Providence, RI July 27-29 (Wakefield et al., 2010a).

Topics for further investigation

Our research has shown significant binaural unmasking in a relatively simple test case. We strongly recommend a research and development program that extends the concepts of granularization to more realistic sources and sea-state noise. Under this work, we believe the filter bank and histogram-based deadband thresholding will give way to more specialized types of signal processing structures by taking advantage of what is already known about the underwater acoustic environment. Even under modest efforts at tuning the algorithm through our work at NUWC, we demonstrated improved performance for real-world scenarios after first showing that the algorithm, as published in the open literature, introduces substantial unwanted artifacts into the audio.

REDUCED COMPUTATIONAL COMPLEXITY OF LARGE MULTI-INPUT SPATIAL AUDIO SYSTEMS

Methods

One of challenges of implementing spatial audio for a large number of directions is the sheer amount of computation required. Figure 44 below shows a listener positioned within a spherical grid of sources (small spheres). Spatializing any one of these sources depends on both the sampling rate of the audio system and the lengths of the head-related impulse responses (HRIR) of the left and right ears. For a sampling rate (f_s) of 44,100 samples/sec and a length (L) of 256 samples for each HRIR, the number of multiplications ($2Lf_s$) is 22.5 million/sec. To render a single source at this rate puts little demand on the processor; to render 60 or more sources becomes burdensome. For this reason, specialized signal processing hardware and software systems have been built to execute brute-force implementations of spatial audio as the number of sources becomes large.

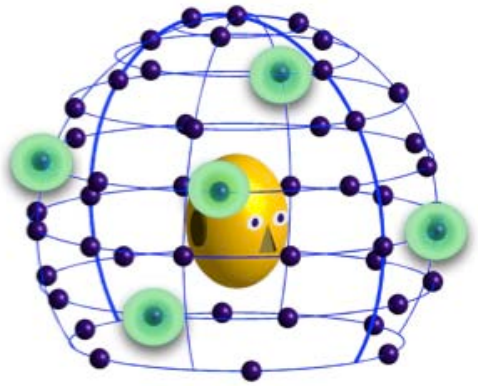

Figure 44. Spatial audio rendering systems must be capable of placing sounds not only at locations for which the listener's HRTF is known, but at all other locations within the sound field. As the number of sounds becomes large, any rendering system that assigns a separate computational element for each sound will inevitably reach its capacity limit. Using ideas as old as stereophonic reproduction, it may be possible to create the same perceptual effect as such brute-force approaches using far fewer computational elements, as suggested by the green sound spheres.

A related challenge for any spatial audio system is that the HRIRs for the observer are known at a limited number of directions in space, yet the system must be able to render a source at *any* direction. As shown in Figure 44, the HRIRs are measured at a finite number of spatial locations (the small spheres) such that, in principle, the wavefield at the two ears for a source at the measured position can be synthesized exactly. However, when a source isn't located at one of the measured positions, its HRIRs must be constructed using properties of the HRIRs at measured locations. Interpolation among a set of measured HRIRs not only introduces potential artifacts in the quality of the spatial rendering, but may also introduce additional computational load, depending on the algorithm used to interpolate.

Rather than focusing on the *acoustics* of the rendering problem, our research has focused on the *psychoacoustics* of spatial hearing with the hope that, like many problems in perceptual engineering, we can discover design principles that obviate the need for a brute-force engineering solution to the spatial rendering problem. Such an approach has served well in a number of audio designs. A classic case is stereophonic reproduction in which two sources (the left and right loudspeakers) can be appropriately mixed to not only create a single phantom source anywhere between the two speaker locations, but an entire orchestra of sources than span the full solid angle. A more contemporary example is 5.1 reproduction (e.g., surround sound) whereby a dense orchestra of sound sources is heard by the listener over an entire 360 degrees by appropriately mixing the source signals presented over loudspeakers at no more than six locations.

As suggested by Figure 44 above, our interest is in determining the minimum number of rendering elements (green sources) by which we can achieve the same perceptual response as a brute-force solution to the problem of rendering the underwater environment for a large number of locations. If we can make this number small, while

also making the costs of interpolation small, then we have achieved our goal of eliminating one of the bottlenecks in effectively deploying spatial audio systems for underwater environments.

Psychophysical Evaluation of Computational-Reduction Techniques Much research has already been done in the area of perceptual design for low-complexity spatial audio. This research has drawn upon techniques from signal processing, control theory, and statistics to reduce the computational costs of rendering sources in 3D. As is typical of these efforts, computationally low-cost solutions are proposed, which are further refined by psychophysical evaluation. For example, our own early work utilized digital filtering modeling to approximate the HRIRs with fewer computational elements. We determined the number of computational elements empirically by asking listeners to discriminate a single source rendered using the measured HRIRs or the approximated HRIRs (Blommer and Wakefield, 1997). In later work, we applied state-space modeling techniques from control science to simultaneously render more than one source (Adams and Wakefield, 2008). Once again, the parameters of the state-space model were optimized empirically by measuring how well listeners could discriminate an approximate rendering of a collection of sources from the original rendering.

In both of the above studies, listeners were asked to utilize any cue, whether spatial, temporal, or spectral in nature, when making their discrimination. Others have asked their listeners to ignore changes in the temporal or spectral character of the source and to focus only on the spatial properties, such as location or diffuseness (Kulkarni and Colburn, 1998). In general, judgments based on *any* difference lead to more conservative, and therefore, computationally more expensive choices of parameters within any given design technique.

Another factor that influences the choice of parameters is the type of source selected for rendering. Listeners are more tolerant of very low-order designs when comparing dynamically varying musical sources than when comparing short bursts of noise (Adams and Wakefield, 2009). Finally, a third factor is the timing and duration of the observation intervals during which listeners hear and compare the two options. Lengthening the inter-observation interval (IOI) reduces the quality of the acoustic memory trace, which can lead to poorer discrimination. Similarly, observation intervals that are too short, relative to the temporal nature of the source, are likely to obscure the cues that a listener would pick up on over extended periods of listening.

Therefore, we expect that the values of an optimal set of parameters within any given reduction technique will depend on the criteria used to judge, the types of sources presented, and the timing of stimulus presentation. What makes spatial audio for underwater environments unique from most other applications is the presence of noise along any given bearing which is “on continuously”, and the need for any rendering system to leave unaltered the perceptual attributes of the sources. Accordingly, our focus during the second year of this study was to develop a psychophysical procedure and evaluate its performance with respect to the particular nature of the underwater environment.

Experiment An experiment was designed to assess the quality of a set of low-computation designs for a spatial audio display of underwater environments. We were interested in determining the number of processing elements necessary for listeners to hear no difference between a fully-rendered and an approximately-rendered environment as a function of the number of sources, their density, and their locations.

Psychophysical Procedure As with any psychophysical procedure, we were most concerned with the problems of fatigue and inattentiveness that arise when performing a large number of observations in which the variations in attribute may be quite subtle from trial to trial. To address these concerns, we defined a *trial* as the evaluation of each of seven design options for a spatially-fixed set of sources. Source positions were varied from trial-to-trial. One of the seven design options was the full rendering of the environment, while the remaining six ranged from low to moderate orders of complexity. This provided a good deal of variation among designs within a given trial, as well as a method to detect overly strict operators who choose to reject any design, including one that is identical to the original.

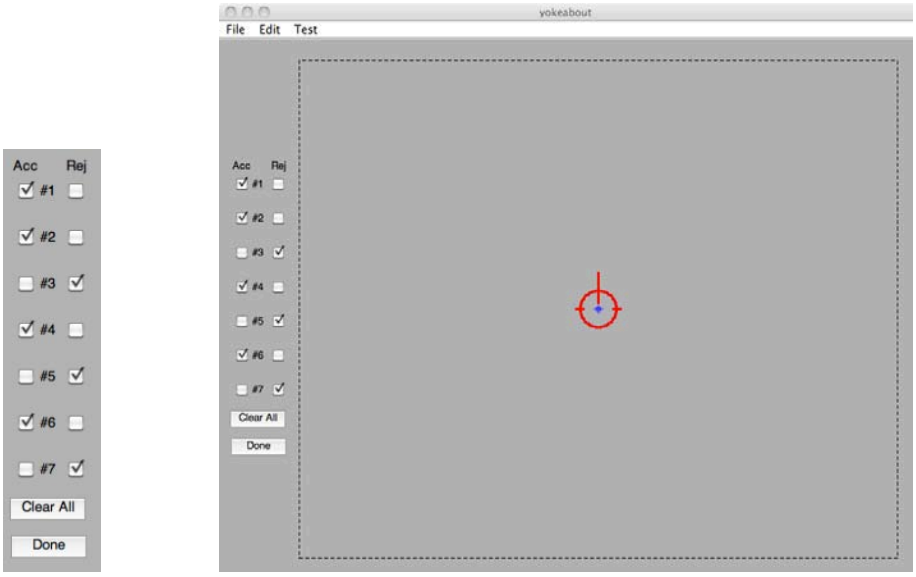

Figure 45. The GUI window for the psychophysical task is shown in the large panel on the right. An icon indicates the position of the listener within the auditory space, as viewed from a top-down perspective. A blow-up of the response panel of check boxes is shown in the left panel. For each design, the user accepts or rejects, and loads their results when the rating task for all seven designs is complete.

The operator used a simple icon displayed within a window on a computer screen to orient to the listening environment. The angular orientation of the icon remained fixed in the forward-looking direction over all trials. Sources were presented at angles along a circle centered at the location of the operator. A selection panel was provided for feedback on the left side of the display window, and is shown above in the larger offset on the left side of Figure 45.

Each trial began with a new source configuration rendered through a full-order system alternating, in time, with the same configuration as rendered through the first design.

Once the operator made their judgment (accept or reject), the second design replaced the first in the alternating sequence. The trial proceeded through each of seven assessments, at which point the operator could submit his or her results. These were stored for further data analysis.

Stimulus timing As shown in the timing diagram in Figure 46, we chose to minimize the IOI inasmuch as sea-state noise in the underwater environment is present continuously in time. We empirically determined that 500-ms observations provided a sufficient period to assess differences between the current design and the full-order system without substantially lengthening the trial.

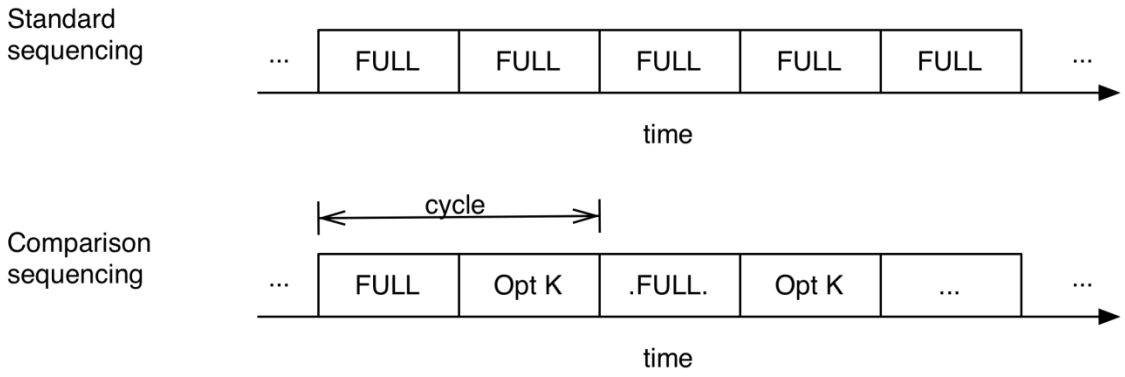

Figure 46. Timing diagram for the spatial rendering of the sources over time. During feedback, the standard sequence uses the full rendering (brute force) approach to construct the sound field. During testing, the current option under consideration alternates with full rendering.

Stimuli Sources were infrapitch pink noise with periods randomly selected between 200 and 300 msec. As the number of sources becomes larger, we have observed that the use of infrapitch pink noise stimuli, in the absence of head motion, tends to keep the overall auditory display from collapsing inside the head.

Source configurations Five different source configurations were used. The single-source condition presented one source during a trial at bearings ranging from -180 to 175 degrees in 5-degree steps. The four-source condition presented four sources, whose positions were drawn a random in 10 degree steps from -180 to 170 degrees. The sixteen-source conditions were broken down into three different spacings. A 1-degree spacing provided a dense cluster of sources isolated to a solid angle of 15 degrees. This solid angle was varied over the circle in steps of 20 degrees. The 5-degree spacing provided a broader angle of activation with lower density. The centroid of the solid angle was sampled every 45 degrees. Finally, random spacing spread the 16 sources over the full circle and drew from locations every 10 degrees.

Trials were blocked according to source configuration. All subjects began with single-source trials and proceeded through four-source, sixteen-source (1 deg. separation), sixteen-source (5 deg. separation), and sixteen-source (scatter).

Full-order rendering used the subject's own HRTFs, as measured at NYU. This rendering system provided samples every 10 degrees from -180 to 170 degrees. For non-measured locations, the HRTFs were interpolated using the desired location's nearest sampled neighbors as follows:

- A minimum-phase section was constructed by computing the mixture of the log magnitude spectra of the nearest neighbors, converting this to linear magnitude, and finding the minimum-phase impulse response for the system.
- A (fractional) group-delay section was constructed by computing the mixture of the group delays of the nearest neighbors, creating the nearest integer-delay impulse response for a 100x oversampled version of the system, and downsampling to the standard sampling rate.
- For both the minimum-phase and group-delay sections, the mixing constant was the difference between the desired bearing and the bearing of one of the neighbors normalized by the difference in bearings of the nearest neighbors (10 degrees, for the NYU measurement system).
- The HRIRs for the interpolated location were formed by convolving the minimum-phase and group-delay subsystems.

Reduction Technique Principal Components Analysis (PCA) was used to reduce the number of computational elements needed for spatializing the sources. Specifically, PCA was performed on the minimum-phase sections of the measured HRTFs and their interpolated HRTFs (in steps of 1 degrees) in the time domain. Weighting coefficients were computed by projecting a given HRIR (either from those measured or those interpolated) onto the span of the principal components. In the experiment, designs of order 3, 5, 7, 9, 11, and 13 were assessed.

Subjects Five music technology students at New York University served as subjects in the experiment. All had audiometrically normal hearing and substantial prior experience in listening to and working with audio signals.

Results

A psychophysical procedure has been designed to aid in the optimization of low-order perceptually lossless computational systems for spatial audio rendering. Results from five subjects show that:

1) The nature and number of sources to be displayed play an extremely important role in optimizing the perceptually lossless system: as the number of sources and the proportion of space they occupy increases, the computational requirements of the system decrease substantially.
2) The results establish that a brute-force technique in which each beam to be spatialized is processed through its own HRTF algorithm is not necessary if the goal is to fully spatialize underwater environments that are noise-dense everywhere.

The top left panel of Figure 47 below provides a sample of the results from one subject. Shown are the results from 72 trials in the One-Source condition. A circle at a given azimuth/design index on this figure indicates that the subject found that design acceptable for an infrapitch pink noise source at that azimuth. As can be seen, one design (#7) was judged acceptable at all bearings, whereas another design (#1) was never judged acceptable.

The top right and bottom left panels present marginal distributions of the data. Though noisy, it appears that azimuths to the left and right ears are the most problematic for the PCA reduction. On average, sources at these bearings were judged acceptable less often than those closer to midline. When averaged across azimuth, there is a clear preference for the higher indexed designs over the lower. The one most preferred (Design #7) was the full-rendering option and that least preferred (Design #1) was the PCA design that used only three principal components to represent all of the locations.

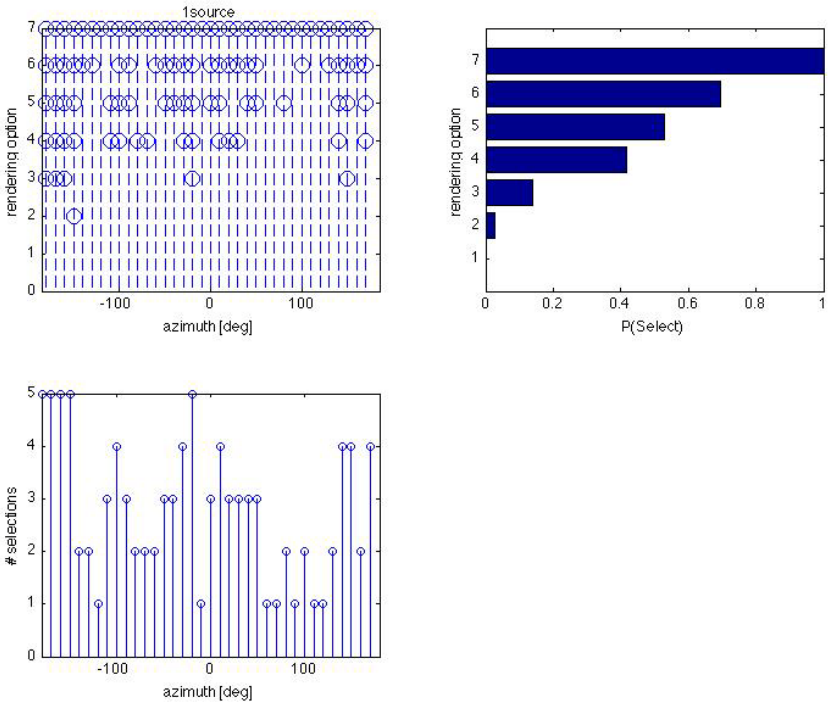

Figure 47. Results from one subject for the single-source condition. The circles displayed in the top left panel indicate that the source located at the given azimuth rendered through the given option was judged to be acceptable. Marginal distributions of the responses in the top left panel are shown in the top right panel (rendering option acceptance, independent of source azimuth) and bottom left panel (azimuth acceptance, independent of rendering option).

Figure 48 below shows the acceptance rates for each design averaged over the five subjects. Standard deviation bars are provided for each data point. As expected, full rendering is highly preferred over all other designs for all stimulus conditions. What is most interesting is that there is a clear dependence of acceptability on source configuration. When listening to a single source, no more than 35% of the locations were

deemed acceptable by listeners, even for a PCA order of 13. In contrast, when listening to sixteen sources scattered randomly over the circle, the average acceptance among listeners is over 98% for a PCA order of 11 and is still above 90% for a PCA order of 9.

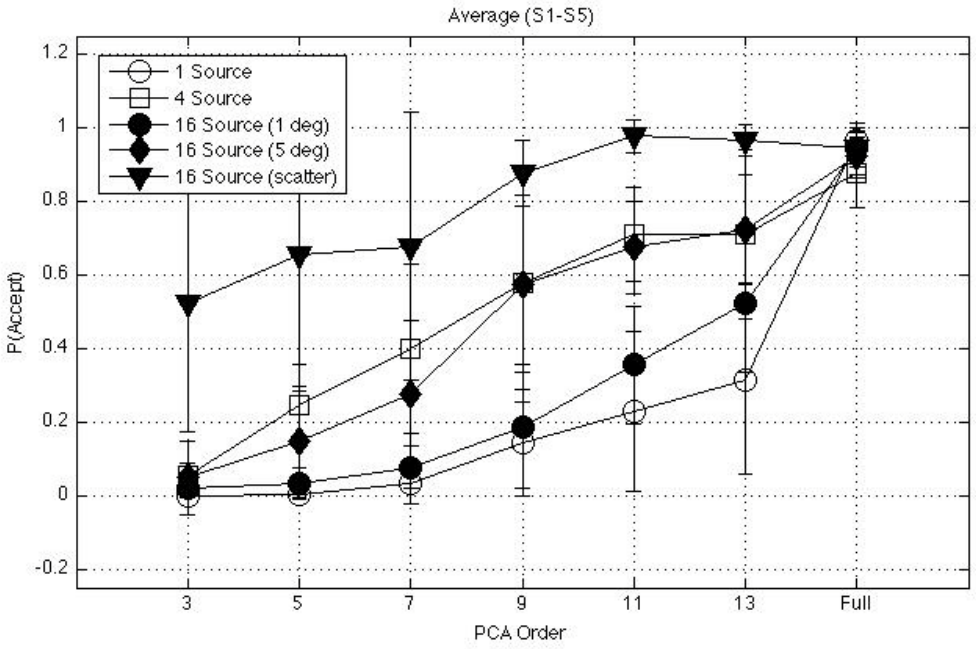

Figure 48. Marginal distribution of rendering option acceptance is shown, averaged over five subjects. The parameter is source configuration. Standard error bars are shown for each datum.

Discussion

Kistler and Wightman (1992) noted that 90% of the variance of the HRTFs could be accounted for by the first 5 principal components. Our results are similar statistically to those of Kistler and Wightman. Nevertheless, even for the case of 13 principal components, where, from a statistical perspective, the relative gain in variance accounted is already vanishingly small, listeners are aware of differences between the fully-rendered and approximately-rendered source. These differences are not necessarily reduced as the number of sources increases. It would appear that whatever is wrong with rendering a single source "remains wrong" when rendering 15 other sources within a small neighborhood. Only when sources are distributed around the full circle does it appear that the quality of rendering improves. In this case, the data show that one would require only half as many processing elements to render 16 sources than using the brute-force approach.

PCA is one among several techniques that share similarities with the computational architecture of a standard full-rendering approach. As such, a good design based on PCA can be easily substituted within any system that has been designed to accomplish the task by brute force.

Summary

We have demonstrated the use of a new psychophysical procedure for rapidly assessing the quality of low-order spatial rendering algorithms. We have seen that as the number of noise sources increases and the proportion of auditory space filled by sources increases, listeners are more tolerant to lower-order spatial rendering algorithms. This is important in light of the computational costs associated with full-rendering, brute-force designs. Results of this study were presented in a publication at the UHSI Symposium in Providence, RI July 27-29 (Wakefield et al., 2010b) and a meeting at the University of Texas, Austin (Wakefield et al., 2010c).

Topics for further investigation

Our research has shown that proportionately less computationally complex systems are required as the number of sources to be rendered increases. For the first time in our discipline, this conclusion is based on a signal-processing system that seamlessly switches among different rendering options, so that a listener's response cannot be contaminated by memory load or other factors.

Our work has also shown that this trend depends strongly on source geometry in the environment. We strongly recommend a research and development program that quantifies performance with respect to a variety of realistic source geometries. This program should also consider more modern signal processing approaches, including that of local basis functions as discussed in the section above. (An important observation is that the original work on local basis functions contains a mathematical flaw, which may actually shed some light on binaural processing if carefully examined.) Any deployed system that relies on sheer computational horsepower will fail to scale when the amount of data and the number of related processing tasks exceed the capacity of the system. To deploy a system that does not employ appropriate and reasonable perceptually-based efficient computational schemes will likely to lead to non-robust, overly sensitive systems; it will certainly lead to a waste of resources.

DUAL TASK DECISION-MAKING WITH SINGLE SENSORY MODALITY INFORMATION SOURCES

Methods

The ability of human sensory-motor and cognitive mechanisms to deal with more than one task at a time has been the subject of extensive investigation and modeling. In this study, the dual task decision-making paradigm from cognitive science was used to explore to what extent the hypothetical "center and surround" auditory and visual channels can function simultaneously and independently. An auditory display composed of independent center and surround sounds representing independent information for unrelated tasks was examined to determine if the two proposed channels could be processed independently and support simultaneous, independent cognitive decision task processes.

The experiment chosen for this study was based on the dual task paradigm used in Schumacher et al., (2001). In those experiments, the stimuli for both tasks were in the same modality (auditory only or visual only) and both were in the "center" fields of their respective modalities. For the visual task, images were presented in the central field of view with a manual button push response required, while for the auditory task, sounds were presented in diotic mode over a headset with a vocal response required.

To investigate the proposed "center/surround" auditory channels in this study, the center auditory task consisted of one of three possible spoken words as might be heard in diotic voice communication over a headset. The words, GO, GET, GUN, were chosen for their similarity and short, single-syllable duration. The listener had to respond by speaking the same word that was heard. The surround task consisted of the presentation of pink noise, an easily localizable broadband acoustic stimulus, at one of three possible locations in the synthetic auditory space around the subject. The three locations were (1) 90 degrees left of the line of sight, (2) 90 degrees right of the line of sight, and (3) directly on the line of sight. The listener was required to respond by pressing one of three buttons in a row on a standard keyboard numeric pad with the right hand (all listeners were right handed). The buttons corresponded logically to the three sound locations; left button for the left sound, right button for the right sound, and center button for the center sound.

The visual-only procedure used in this study was the analog of the auditory-only paradigm described above. A word of text was presented at a central fixation point (the proposed visual “center” channel) and a large diamond-shaped object was presented in the left, right, or upper visual periphery (the “surround” channel). Required responses to these visual stimuli were to speak the presented word and to press the appropriate button for the location of the peripheral object.

Five New York University students performed at least five auditory-only test sessions and at least five visual-only sessions for at total of at least ten hours of testing over about six days for each subject. For the surround, HRTFs built from measurements made on the KEMAR manikin in the anechoic chamber were used to present a pink noise source at the

appropriate spatial location over headphones. For the center, the speech token was presented diotically over headphones.

Each test session consisted of six pure blocks and 10 mixed-trial blocks corresponding to the design used in Schumacher. There were three pure blocks of center stimuli, to which subjects were asked to respond verbally, and three pure blocks of surround stimuli, which subjects used a manual response mechanism involving a computer keyboard. There were a total of 45 stimuli presented in each pure block. Mixed blocks consisted of a total of 48 randomly presented trials including 15 center-only stimuli, 15 surround-only stimuli and 18 center and surround stimuli presented concurrently. Each session took subjects approximately 1 hour to complete. Subjects were asked to participate in two sessions per day. The first session was used as training and the results were eliminated from analysis.

Instructions to the subjects stressed the requirements for both speedy and accurate dual task performance. No explicit identification of either task as primary or secondary was given in the instructions. They were told to attend to both tasks equally as they occurred. Automatic feedback was given in the pure manual response blocks and errors in the pure verbal response blocks were identified by the test monitor as they occurred.

Results

The dual task decision-making paradigm from cognitive science was used to determine to what extent the proposed "center and surround" auditory and visual channels can function simultaneously and independently. Auditory and visual displays composed of independent center and surround sources representing independent information for unrelated tasks were tested to determine if the two proposed channels in each modality can be processed independently and support simultaneous, independent cognitive decision task processes.
Results show that:

1) Interference exists when two tasks are presented simultaneously in the auditory or visual modality, but
2) Time-sharing occurs:
    a. Auditory dual task time increases by ~20% when two tasks are presented in center-surround mode
    b. Visual dual task time increases by 20-30% when tasks are presented in center-surround.

These findings of faster response and less interference in the auditory modality recommend consideration of this sensory modality instead of (or in addition to) vision when designing the human interface for decision support systems.

Figure 49 through Figure 55 show results of the mean response times for the auditory-only experiment. Results are shown for each session for the following responses:
−○− Manual: Pure Manual responses to “surround” stimuli

–●– Manual Heterogeneous: In Mixed Verbal/Manual session, responses to stimuli presented using the "surround" mode only
–●– Manual Mixed: In Mixed Verbal/Manual session, Manual responses to "center" and "surround" stimuli presented simultaneously
–□– Verbal: Pure Verbal responses to "center" stimuli
–■– Verbal Heterogeneous: In Mixed Verbal/Manual session, responses to stimuli presented using the "center" mode only
–■– Verbal Mixed: In Mixed Verbal/Manual session, Verbal responses to "center" and "surround" stimuli presented simultaneously

It can be observed from Figure 49 to Figure 55 that, although some subjects show an improvement in performance with each session indicating effect of learning, in most cases training did not contribute to a decrease in response times.

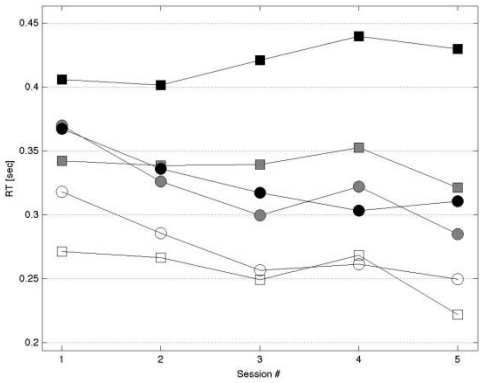

Figure 49. Subject 1 mean response times per session, for dual auditory experiment. See text for legend

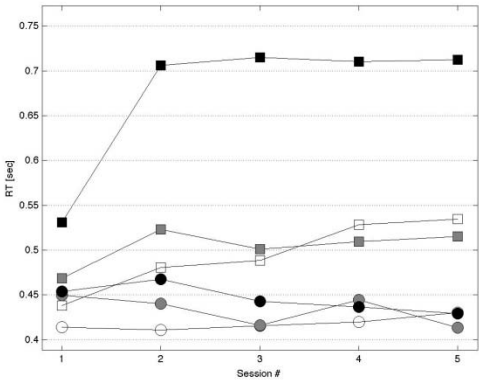

Figure 51. Subject 3 mean response times per session, for dual auditory experiment. See text for legend

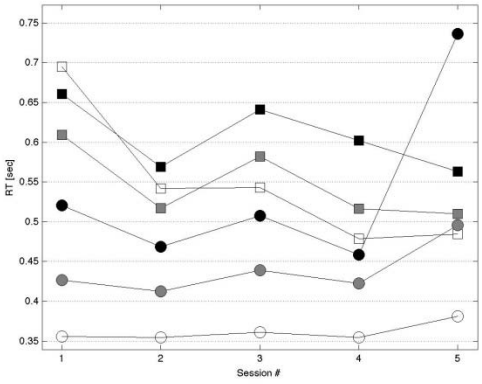

Figure 50. Subject 2 mean response times per session, for dual auditory experiment. See text for legend

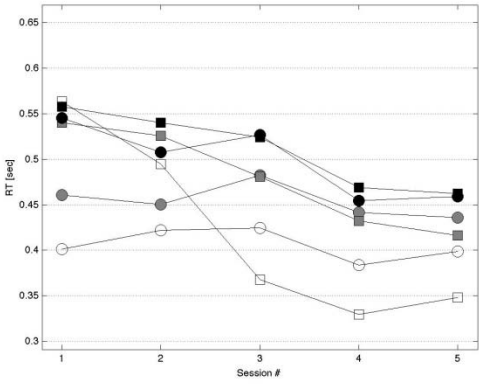

Figure 52. Subject 4 mean response times per session, for dual auditory experiment. See text for legend

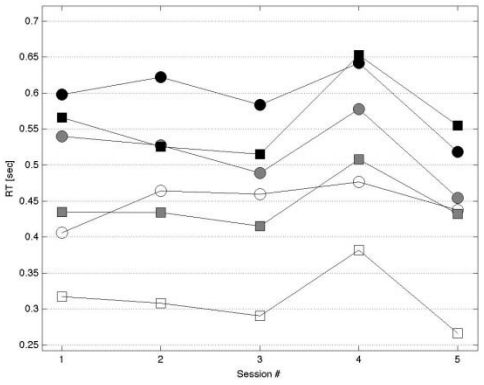

Figure 53. Subject 5 mean response times per session, for dual auditory experiment. See text for legend

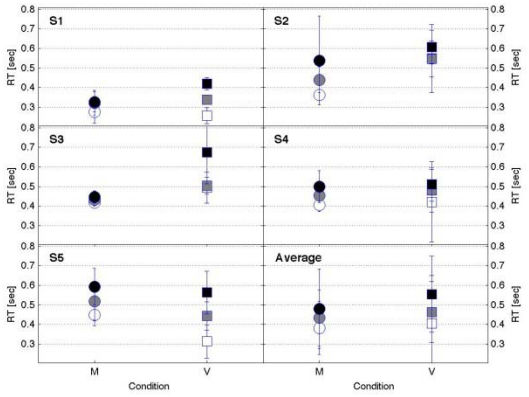

Figure 54. Average results for dual auditory experiment for each subject, and mean across all subjects for sessions 2-5. See text for legend

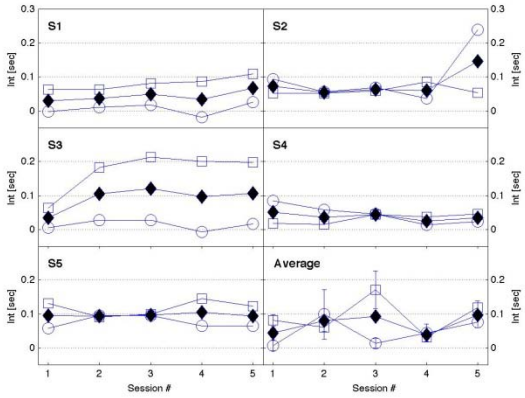

Figure 55. Interference times for dual auditory task for center stimuli/verbal responses (square), surround stimuli/manual responses (circle), average interference time (diamond).

Interference times (shown in Figure 55) are determined by the difference in the response times in the Mixed Verbal/Manual tasks, between the heterogeneous and dual trials. In other words, it is the difference in response times between when center and surround stimuli are presented concurrently, and when center *or* surround stimuli are presented in isolation, in a mixed session. This difference in response time is the interference of the dual modality. Some subjects (S1 and S3) show a consistent and significant difference in interference times between the surround and center stimuli.

It can be observed that the best performance times are obtained in the pure tasks – where manual-only or verbal-only responses are required and expected by the subject. Subjects' response times increased in the dual tasks, by an average of 51 ms for the heterogeneous stimuli. The increase is, in some cases, up to 200 ms. The average interference between the dual heterogeneous and the dual trials (where the center and surround stimuli were presented concurrently) is 90 ms.

Dual Visual Task Subject response times for the visual dual task for each session are presented in Figure 56 through Figure 61. Similarly to the dual auditory experiment described above, the effect of learning does not have a significant impact on response times.

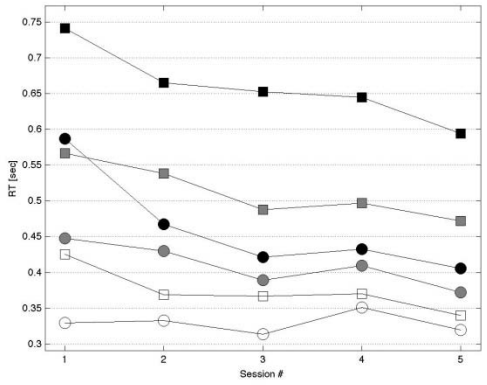

Figure 56. Subject 1 mean response times per session, for dual visual experiment. See text for legend.

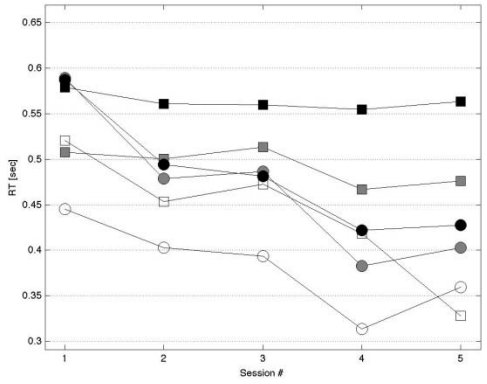

Figure 57. Subject 2 mean response times per session, for dual visual experiment. See text for legend.

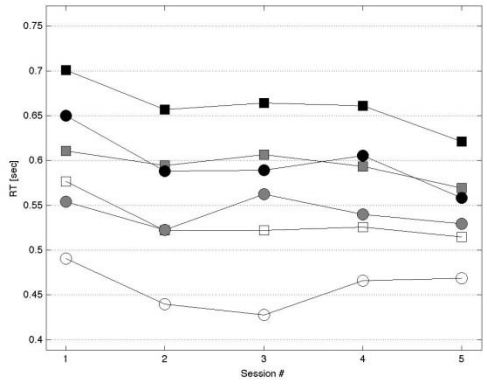

Figure 58. Subject 3 mean response times per session, for dual visual experiment. See text for legend.

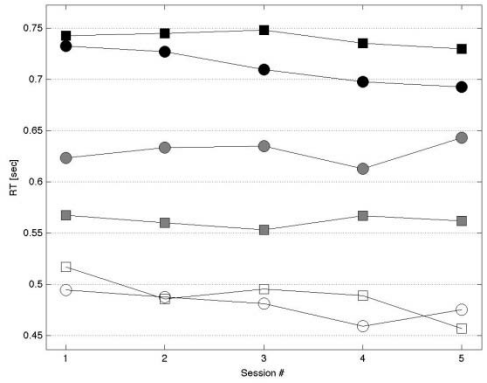

Figure 59. Subject 4 mean response times per session, for dual visual experiment. See text for legend.

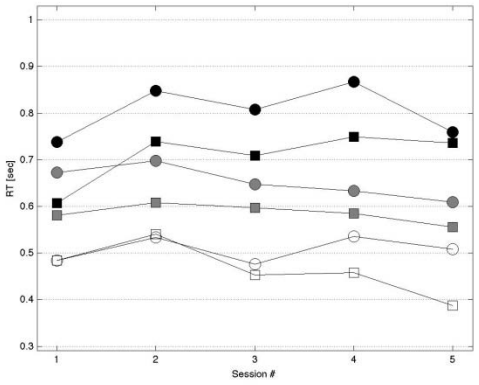

Figure 60. Subject 5 mean response times per session, for dual visual experiment. See text for legend.

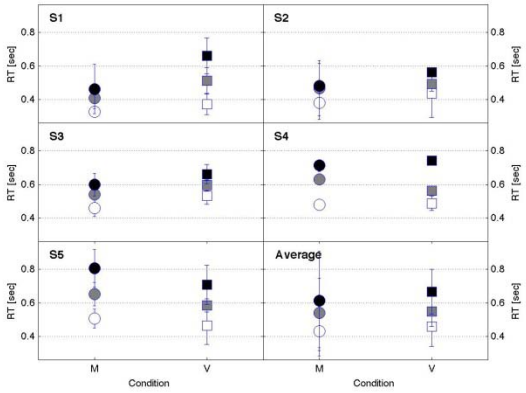

Figure 61. Average results for dual visual experiment for each subject, and average performance across all subjects for sessions 2-5. See text for legend.

The mean response times for each subject for all sessions, as well as the mean responses for all subjects for the dual visual task are presented in Figure 61. This figure shows the response time for the Manual and Verbal Pure (white), Heterogeneous (grey), and Dual (black) trials.

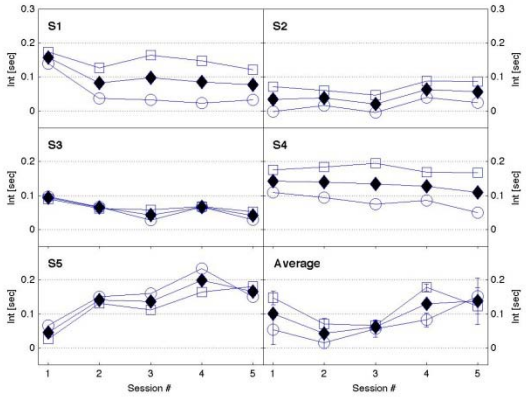

Figure 62. Interference times for dual visual task for center stimuli /verbal responses (square), surround stimuli/manual responses (circle), and average interference time (diamond). Times are presented for each subject, and the mean for all subjects per session.

Interference times in the mixed trials are presented in Figure 62. It can be observed that three subjects (S1, S2 and S4) had lower interference times for the surround stimuli (with manual responses) in each session, with a difference of approximately 100 ms in interference times, while the two other subjects (S3 and S5) had overall similar interference times in the surround and center stimuli cases. It is important to note that the difference in interference times between subjects is not the same in the dual auditory task, indicating that the behavior in the visual and auditory tasks for each subject is different. The average interference times in the dual visual task between the pure and heterogeneous stimuli is 109 ms, for all sessions and all subjects. The average interference times between the heterogeneous and dual stimuli is 94 ms.

Auditory vs. Visual Interference The dual center-surround interference behavior is different in the auditory and visual modalities. A comparison of the interference between the auditory and visual tasks are presented in Figure 63. The figure shows the mean interference times for individual subjects for manual responses/surround stimuli (open symbols) and for verbal responses/center stimuli (closed symbols) under auditory presentation (abscissa) vs. visual presentation (ordinate). Each symbol represents responses for one of the five subjects: S1 (left triangle), S2 (right triangle), S3 (diamond), S4 (up triangle), S5 (down triangle).

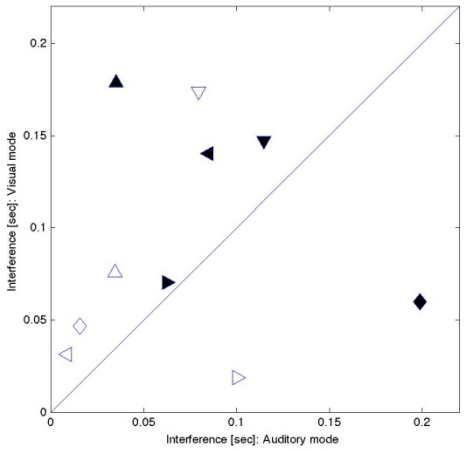

Figure 63. Scatter plot of mean interference times for 5 subjects. Responses are shown for manual (open symbols), and verbal (closed symbols) responses.

Discussion

Averaged reaction times over all subjects are shown in Figure 64 where dual and heterogeneous responses are separated into auditory-only and visual-only presentation modes. The data extend the Schumacher finding that, for certain individuals, simple cognitive decisions can be performed in "virtually perfect" time sharing; i.e., mean interference times of 10 - 20 ms between dual and single task response in the condition of equally likely dual or single task trials. Unlike Schumacher, the present data are for center-surround stimulus organizations in auditory-only and visual-only modalities. While the mean interference times for these single modality organizations are greater than those found by Schumacher using dual modalities, they still amount to only a 20% delay in the auditory-only mode and a 30% delay for the visual-only mode. Hence, for this organization, two independent stimuli can be perceived almost simultaneously by each pathway and processed by decision-making centers with only minor time delay to either decision process. The decision-maker still gains a significant time advantage doing the tasks together rather than doing them one at a time.

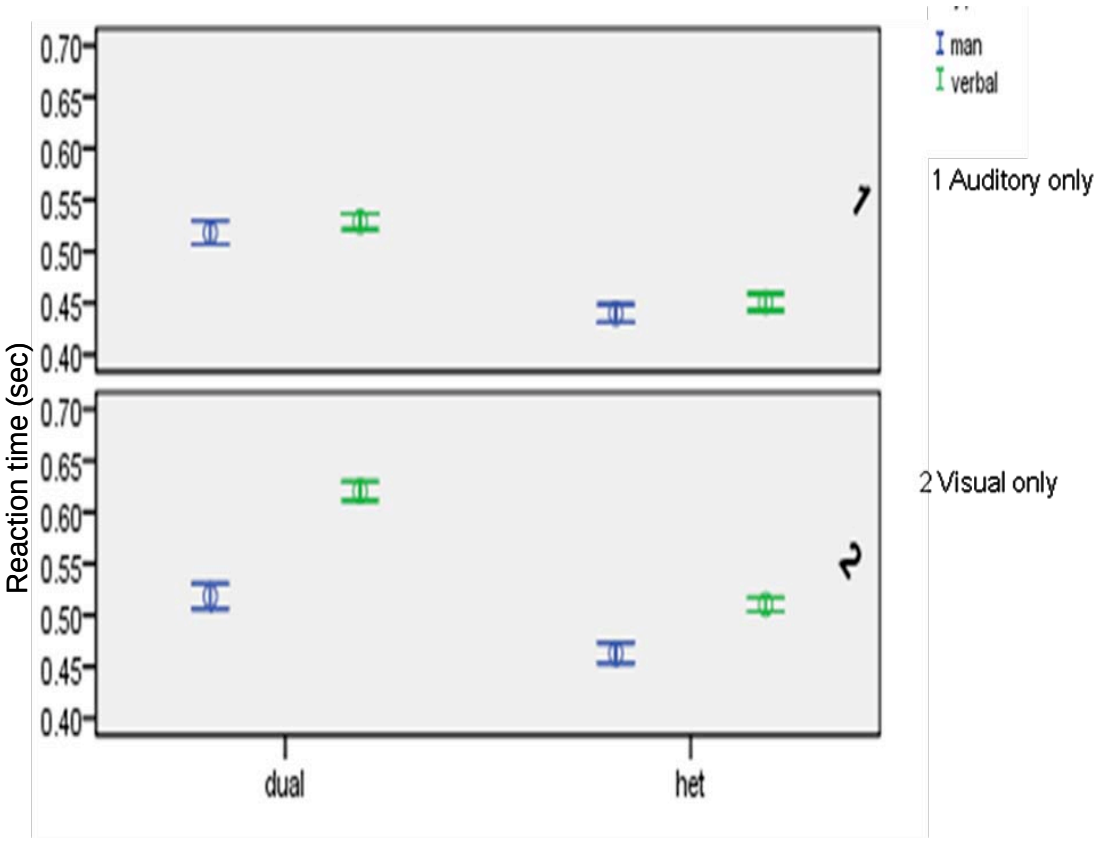

Figure 64 Average reaction times for 5 subjects. Responses are shown for manual (blue symbols), and verbal (green symbols) responses in dual and heterogeneous modes for auditory-only and visual-only presentations.

The data also show that a manual button press response can be done faster than a word verbalization response and that, more interesting, either response to auditory stimuli can be expected to be faster than the corresponding response to visual stimuli, whether center or surround in nature.

Individual capabilities for low interference delay processing vary across subjects according to modality. For example, in the verbal response task, Subject 4 has much lower interference with the auditory-only modality than the visual-only while subject 3 has the opposite tendency. However, over all subjects the trend, as shown in Figure 64, heavily favors less interference in the auditory modality than the visual. These findings of faster response and less interference in the auditory modality recommend consideration of this sensory modality instead of or in addition to vision when designing the human interface for decision support systems.

The experiments on dual task decision-making were presented in a publication at the 127thAudio Engineering Society Convention; October 9-12, in New York (Santoro et al., 2009) and at the UHSI Symposium in Providence, RI July 27-29 (Santoro et al., 2010).

REFERENCES

- Adams, N. H. and Wakefield, G. H. (2009). "State-Space Models of Head-Related Transfer Functions for Virtual Auditory Scene Synthesis," J. Acoust. Soc. Am., 125(6), 3894-3902.
- Adams, N. H. and Wakefield, G. H. (2008). "State-Space Synthesis of Virtual Auditory Space," IEEE Transactions on Audio, Speech and Language Processing, 16(5), 881-890.
- Begault, D. R., Wenzel, E.M., Anderson, M. (2001). Direct Comparison of the Impact of Head Tracking Reverberation, and Individualized Head-Related Transfer Functions on the Spatial Perception of a Virtual Speech Source, J. Audio Eng. Soc. 49(10), 904–917.
- Begault, D.R., Godfroy, M., Miller, J.D., Roginska, A., Anderson, M.R., and Wenzel, E.M. (2006) Design and Verification of HeadZap, a Semi-Automated HRIR Measurement System. 120th Convention Audio Engineering Society, Paris, France, 20-23 May 2006.
- Blommer, M. and Wakefield, G. H. (1997). "Pole-zero Approximations For Head-related Transfer Functions Using A Logarithmic Error Criterion," IEEE Trans. Speech and Audio Processing, Vol. 5, 278-287.
- Brungart, D. S., Chang, P. S., Simpson, B. D., and Wang, D. (2009). “Multitalker speech perception with ideal time-frequency segregation: Effects of voice characteristics and number of talkers,” J. Acoust. Soc. Am. 125, 4006-22.
- Brungart, D. S., and Simpson, B. D. (2002). “The effects of spatial separation in distance on the informational and energetic masking of a nearby speech signal,” J. Acoust. Soc. Am. 112, 664–676.
- Brungart, D. (2001). “Informational and energetic masking effects in the perception of two simultaneous talkers,” J. Acoust. Soc. Am., 109, 1101-09.
- Buechner S. J., Hölscher, C. & Wiener, J., (2009) “Search Strategies and their Success in a Virtual Maze”. Proceedings of the 31th Annual Conference of the Cognitive Science Society, 1066-1071.
- Gabor, D. (1946). "Theory of communication," J. IEE (London), 93(III), 429-457.
- Hill E.W, Rieser J.J., Hill M.M., Halpin J., (1993) “How persons with visual impairments explore novel spaces: strategies of good and poor performers.” Journal of Visual Impairment and Blindness, 87(8).
- Kistler, D. J. and Wightman, F. L. (1992). “A model of head-related transfer functions based on principal components analysis and minimum-phase reconstruction ,” J. Acoust. Soc. Am., 91, 1637-1647.
- Kulkarni, A., and Colburn, H. S. (1998). “Role of spectral detail in sound- source localization,” Nature,London, 396, 747–749.

- Levitt, H. (1971). “Transformed up-down methods in psychoacoustics,” J. Acoust. Soc. Am., 49, 467-477.
- Roginska, A., Wakefield, G.H., Santoro, T.P. (2010a). "Stimulus-dependent HRTF preference". Proceedings of the 129th Audio Engineering Society Convention, November 4-7, San Francisco, CA.
- Roginska, A., Wakefield, G.H., Santoro, T.S. (2010b) “Use of Interfaces in an Auditory Environment during a Navigation and Search Task”, Accepted for publication in the proceedings of the UHSI Symposium, July 27-29, Providence, RI.
- Roginska, A., Wakefield, G.H., Santoro, T.S. (2010c) “User Selected HRTFs: Reduced Complexity and Improved Perception”, Accepted for publication in the proceedings of the UHSI Symposium, July 27-29, Providence, RI.
- Roginska, A., Wakefield, G.H., Santoro, T.P., and McMullen, K. (2010d). "Effects of interface type on navigation in a virtual spatial auditory environment". Proceedings of the 16th International Conference on Auditory Displays, June 9-15, Washington, DC.
- Santoro, T.S., Wakefield, G.H., Roginska, A. (2010) “Investigations of Binaural Listening HSI for Sonar”, UHSI Symposium, July 27-29, Providence, RI
- Santoro, T.P., Roginska, A., Wakefield, G.H. (2009) "Listening within you and without you: center-surround listening in multi-modal displays". Proceedings of the 127thAudio Engineering Society Convention; 2009 October 9-12, New York, NY.
- Santoro, T.P. and Wakefield, G.H. (2005). Spatialized auditory displays for passive sonar listening. NSMRL Technical Report TR1233, April 26, 2005.
- Schumacher, E.H., Seymour, T.L., Glass, J.M., Fencsik, D., Lauber, E.J., Kieras, D.D., & Meyer, D.D. (2001) Virtually perfect time-sharing in dual-task performance: Uncorking the central cognitive bottleneck. *Psychological Science*, 2001, 12, 101-108.
- Thinus-Blanc, C. & Gaunet, F., (1997) “Representation of space in blind persons: Vision as a spatial sense?”. Psychological Bulletin, 121, 20-42.
- Wakefield, G.H., Roginska, A., Santoro, T.S. (2010a) “Enhancement of Perceived Figure-Ground for Spatial Audio in Underwater Environments”, Accepted for publication in the proceedings of the UHSI Symposium, July 27-29, Providence, RI.
- Wakefield, G.H., Roginska, A., Santoro, T.S. (2010b) “Human-Constrained Design of Spatial Audio Systems Accepted for publication in the proceedings of the UHSI Symposium, July 27-29, Providence, RI.
- Wakefield, G.H., Roginska, A., Santoro, T.S. (2010c) "State-of-the-Art Multi-Modal Tools for Increased Situational Awareness in Sonar Applied Binaural Research" Brief to the Signal Processing Working Group Meeting at University of Texas, Austin, February 11, 2010.
- Warren, R.M., & Bashford, J.A., Jr. (1981) “Perception of acoustic iterance: Pitch and infrapitch”. Perception & Psychophysics, 29, 395-402.